Institut für Maschinelle Sprachverarbeitung
Universität Stuttgart
Pfaffenwaldring 5B
D-70569 Stuttgart

Master Thesis

Content-Aware Text-To-Speech with Prompt-Based Prosody Control

Thomas Bott

Studiengang:	M.Sc. Computational Linguistics

Prüfer:	Prof. Dr. Thang Vu
	Dr. Antje Schweitzer
Betreuer:	Florian Lux

Beginn der Arbeit:	15.02.2023
Ende der Arbeit:	15.08.2023

Erklärung (Statement of Authorship)

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbstständig verfasst habe und dabei keine andere als die angegebene Literatur verwendet habe. Alle Zitate und sinngemäßen Entlehnungen sind als solche unter genauer Angabe der Quelle gekennzeichnet. Die eingereichte Arbeit ist weder vollständig noch in wesentlichen Teilen Gegenstand eines anderen Prüfungsverfahrens gewesen. Sie ist weder vollständig noch in Teilen bereits veröffentlicht. Die beigefügte elektronische Version stimmt mit dem Druckexemplar überein.

(Thomas Bott)

Abstract

This thesis proposes a text-to-speech system that is conditioned on sentences embeddings extracted from natural language prompts in order to make the prosodic parameters of generated speech controllable in an intuitive and effective way. The system builds on a transformer-based TTS architecture and provides benefits regarding speed, data efficiency, robustness and controllability. The proposed integration scheme essentially concatenates speaker and sentence embeddings by modeling inter-dependencies between them before inducing the joint representation into the model. Furthermore, a training strategy is developed that operates on merged emotional speech and text datasets and varies prompts in each iteration, increasing the generalization capabilities of the model and reducing the risk of over-fitting. Extensive objective and subjective evaluations on utterances generated from sentences of emotional text datasets demonstrate the prompting capabilities of the conditioned TTS system. It achieves high prosodic controllability whereby the emotional content of provided prompts is transferred accurately to generated speech. At the same time the system maintains precise tractability of speaker identities as well as overall high speech quality and intelligibility. Besides a high correlation between prompts and speech prosody, fine-tuning the sentence embedding extractor has been found to be crucial. The proposed TTS system is limited with regard to modeling unseen speakers, intensities and multiple languages.

Contents

List of Tables

List of Figures

1 Introduction

With the advances in machine learning during the past years, text-to-speech systems have become very powerful. In some applications they almost reach quality comparable to human speakers in terms of perceived naturalness (Ren et al., 2020). One of the main difficulties in speech synthesis is the one to many mapping problem which describes the fact that for a given input text there are many possible spectrogram representations which can differ in various components such as speaking style, intonation, stress or rhythm, collectively referred to as prosody. It remains a challenge to generate a spectrogram representation with appropriate and natural sounding prosody especially with respect to the content of the input or its context. Therefore, in recent years a number of different approaches have been proposed to tackle this challenge.

Modern transformer-based text-to-speech synthesis systems introduce a component which converts the input text into phonemes before encoding them. This comes with the benefit that no text preprocessing is necessary which can be complicated and expensive. However, by doing so, information about words, sentence structure and meaning of the input text is lost. Such information has been found to be closely related to the prosodic properties of speech (Friederici, 2001; Baumann and Riester, 2012). Numerous approaches focus on providing syntactic information to the speech synthesis model which can be helpful for deciding which words should be stressed or where intonation boundaries are located (Hirschberg and Rambow, 2001; Mishra et al., 2015; Köhn et al., 2018). Other work tries to induce the meaning of the words or even the whole input text to the model which can also be useful to determine an appropriate prosody by providing information about the desired sentiment or emotion in which the text should be synthesized as well as syntactic information to a certain degree (Hayashi et al., 2019). Thereby, the assumption is that there exists a correlation between the meaning of the utterance and its prosodic parameters.

This thesis builds on the idea of providing the meaning of the input text to the speech synthesis model and introduces a method which allows to control the prosody of generated speech with natural language prompts. The main idea comes down to additionally conditioning the model on embeddings obtained by a large language model during training. At inference the additional information can be helpful for generating a fitting prosody with respect to the provided prompt. Prompting methods have been applied to text and image generation with great success to guide the model to generate the desired output (e.g. Ramesh et al. (2022)). However so far there is only a small amount of research on prompt-based text-to-speech synthesis.

Previous work on controllable TTS, which uses style factors such as prosodic tags (Guo et al., 2022a) or pitch contours (Bak et al., 2021), require specific values at inference that are often only understandable with acoustic knowledge. TTS systems that infer the desired prosody from reference speech (Wang et al., 2018), oblige users to provide appropriate speech samples, which can be time-consuming and difficult to obtain. In this vein the proposed approach in this thesis presents a simpler and more intuitive way to control the prosodic parameters of synthesized speech since it depends on prompts in natural language form.

The proposed system builds on a baseline which combines architectural designs from several TTS models proposed in previous work and provides benefits regarding speed, data efficiency, robustness and controllability (Lux et al., 2023a). The proposed integration scheme concatenates speaker and sentence embeddings by modeling inter-dependencies between them before inducing the joint representation into the model's encoder, decoder and prosody predictors. This approach is simple and adds only little computational cost in comparison to the baseline. Furthermore, a training strategy is developed that operates on merged emotional speech and text datasets and varies prompts in each iteration, potentially increasing the generalization capabilities of the model while at the same time reducing the risk of over-fitting. In addition, combining existing datasets eliminates the necessity of manually providing style descriptions along with the speech data. Objective and subjective evaluations are performed in order to assess the controllability of generated prosody through provided prompts.

Thus, the main research question can be formulated as follows:
Is it possible to accurately control the prosodic parameters of speech generated by a text-to-speech system through providing natural language prompts?

Thereby, the hypotheses are:

1. Conditioning a speech synthesis model on sentence embeddings of textual prompts improves the quality of generated speech in terms of diverse prosody and naturalness.
2. The prosodic parameters of speech and the conveyed emotional state are controllable by providing suitable prompts during inference.
3. The emotional content of the prompts is correlated to the prosody and conveyed emotion of synthesized speech.
4. Semantically similar prompts result in similar prosody and conveyed emotion.
5. The controllability of speaker identities is not affected by the integration of sentence embeddings.

This thesis begins with a description of the fundamental background (section 2) followed by an overview of related work focusing TTS systems incorporating supplementary information (section 3). Further, the baseline along with the extension of the proposed system are explained in detail (section 4). Section 5 outlines the conducted experiments and applied objective as well as subjective evaluation methods. Subsequently, the collected results are presented (section 6) and discussed (section 7). Finally, the limitations of the proposed approach and possible future research directions are pointed out (section 8). The most important findings of the thesis are summarized in section 9.
The implementation of the TTS systems and trained models can be accessed on GitHub [1]. Furthermore, a demo page containing several audio samples is provided [2] along with an interactive space on HuggingFace [3].

[1] `https://github.com/Thommy96/IMS-Toucan`

[2] `https://thommy96.github.io/toucanpromptingdemo/`

[3] `https://huggingface.co/spaces/Thommy96/promptingtoucan`

2 Background

2.1 Deep Learning Methods

Deep Learning is a branch of machine learning that plays a central role in today's state-of-the-art algorithms and artificial intelligence models and is very powerful in processing and learning from large amounts of data. It has been applied with great success in many areas like healthcare (Ma et al., 2015; Xiong et al., 2015), image generation (Krizhevsky et al., 2012), natural language processing tasks (Collobert et al., 2011; Bordes et al., 2014; Sutskever et al., 2014), speech recognition (Hinton et al., 2012) and speech synthesis (Wang et al., 2017; Ren et al., 2020).

Deep Learning builds on the idea of neural networks which are designed to model the way human brains process information (McCulloch and Pitts, 1943; Rosenblatt, 1958). Neural networks typically comprise interconnected layers each consisting of a number of processing units called neurons. The neurons and their connections possess individual weights and bias values which collectively form the parameters of the neural network. During the training process these parameters are adjusted in a way such that the network adapts to the training data.

Deep learning gets its name from the fact that it creates a deep structure by stacking additional layers. Thereby, the idea is that layers at the beginning learn simple features while deeper layers specialize in comprehending complex attributes. Such a network can also be seen as a powerful hierarchical feature representation of the input data.

2.1.1 Feed Forward Neural Networks

In a feed forward neural network the information is processed only in one direction, i.e. forward through the layers of the network. In its simplest configuration it consists of a single layer. In this case the values of the input vector X are multiplied by the weights W. Thereby W is a matrix of shape $|X| \times N$ where N corresponds to the number of neurons since each input value has a connection to each neuron. The weighted values are then subsequently aggregated along with the biases $B = \{b_1, ...b_N\}$ to form the weighted sum Z. This sum is then applied to an activation function σ in order to ensure that the output $Y = \{y_1, ..., y_N\}$ is mapped between required values such as $(0, 1)$ or $(-1, 1)$.

Importantly, due to the use of matrix multiplications in the calculation, the input X can also be a matrix encompassing multiple input vectors. The calculation of the output Y is presented in equation 1.

$$Y = \sigma(WX + B) \tag{1}$$

The major limitation of single layer neural networks is that they can only handle linearly separable data. Introducing more than one layer to the network solves this shortcoming and therefore makes it more powerful. The additional layers between the input and the output layer are called hidden layers. In the calculation of Y the input of each layer is the output of the previous layer. Since there are multiple layers, a weight matrix W^l and a bias vector B^l for each layer l are required. Equation 1 can now be updated as in equation 2 which describes the so called forward propagation of a neural network.

$$Y = \sigma(W^L ... \sigma(W^2 \sigma(W^1 X + B^1) + B^2) ... + B^L) \tag{2}$$

The purpose of having activation functions after each layer is to introduce non-linearity to the network. Commonly used activation functions are *sigmoid*, *tanh* or *ReLU*. If the output should be a probability distribution, a softmax function can be applied to the final layer.

In order to fit the neural network to the training data its parameters have to be updated. After the forward propagation, a loss function is applied which compares the output with the target training data. The overall goal is to adjust the parameters of the network in such a way that the loss becomes as small as possible. This is achieved by backpropagating the loss through the layers of the network. Thereby the gradients of the loss with respect to the weights and biases at each layer are calculated by taking the partial derivative of the loss function. The gradients describe the direction and magnitude of the changes required to reduce the loss. Then the parameters are adjusted using an optimization algorithm like gradient descent or Adam which makes use of an adaptive learning rate and momentum to improve convergence and performance (Kingma and Ba, 2014).

2.1.2 Convolutional Neural Networks

Convolutional Neural Networks (CNNs) contain at least one convolutional layer followed by a pooling layer and a fully connected layer as final component. In the convolutional layer a filter matrix, also known as kernel, is shifted across the values of the input data, applying a dot product at each step. In this way the input matrix is folded with the goal of detecting local features that are important for the task at hand. The values of the filter are trainable parameters and it is also possible to apply multiple different filters each focusing on distinct feature aspects of the input. Furthermore multiple convolutional layers can be stacked in order to increase the complexity of the network.

As in feed forward neural networks an activation function is applied after each convolutional layer which introduces non-linearity. The output of a convolution is processed by a pooling layer that conducts a dimensionality reduction by sliding a filter over the values and applying an aggregation function. The two main types of pooling layers are max pooling where the filter selects the maximum value at each step and average pooling where the values within the filter range are averaged. The motivation behind the pooling operation is to maintain only the most important features and thus improve the efficiency of the network. The final layer of a convolutional neural network is a fully connected layer with e.g. softmax activation in order to output a probability distribution.

2.1.3 Recurrent Neural Networks

Recurrent Neural Networks were popularized by Hopfield (1982) and are useful to process sequential data, e.g. a series of data points over time or text. The major difference to feed forward or convolutional neural networks is that they take information from previous inputs into account which influences the current processing step. Therefore, the output after treating the last element of the input sequence is dependent on all previous elements. If the network is bidirectional it even has information about future elements at each time step. This is achieved by sharing a memory matrix of weights across the network's layers which is fed into the calculation at each processing step and then updated accordingly.

Recurrent Neural Networks tend to have problems with exploding or vanishing gradients. These describe the issue that during backpropagation the gradients become increasingly large or small as a consequence of the memory updates. A popular architecture of recurrent neural networks is the Long short-term memory (LSTM) which was introduced by Hochreiter and Schmidhuber (1997). This architecture addresses the vanishing gradient problem by accounting for long-term dependencies in the input. For later parts of the input, traditional recurrent neural networks already have forgotten about distant parts at the beginning of the sequence since the memory has been updated too many times in between. To tackle this issue the neurons in a LSTM are cells that can control which information should be stored and what shall be forgotten when updating the memory. This is achieved by a forget gate, an input gate and an output gate, each only letting through specific information based on the output of the previous time step and the input of the current time step. Like that important information from the the beginning of the sequence can be stored throughout all processing steps. The Gated Recurrent Unit (GRU), introduced by Cho et al. (2014) simplifies the architecture of LSTM by merging the forget and input gates, making it computationally more efficient while still capturing long-term dependencies.

2.1.4 Sequence to Sequence Learning

In Sequence to Sequence Learning a recurrent neural network is trained to map the input sequence to an output sequence as in e.g. machine translation, speech recognition or speech synthesis. The architectures of such approaches comprise an encoder and a decoder recurrent neural network. The encoder processes the input sequence and stores the information in a hidden state which is then passed to the decoder whose task is to predict the target sequence.

A very important paradigm which improves the performance of sequence to sequence learning models is the attention mechanism. It was introduced in Bahdanau et al. (2014) and addresses the bottleneck problem of only having a single hidden state as input for the decoder which limits the amount of information the decoder receives, especially for long sequences. The attention mechanism gives the decoder access to not only the last, but all hidden states of the encoder such that at each decoding time step the most relevant information can be used. This enables the encoder to pay attention to different parts of the input at each step. Thereby a weighted sum of the encoder states with learned attention weights is computed, indicating which state should be used for each decoding step.

The paper *Attention is All You Need* by Vaswani et al. (2017) takes this idea one step further and introduces the transformer model which is based solely on attention mechanisms without the need for recurrent neural networks. Instead the encoder and decoder blocks use self-attention to process information. In contrast to the attention mechanism between decoder and encoder, self-attention operates among states of the same block, e.g. all encoder states. The intuition stays the same as before: for each hidden state of the sequence the most relevant information of the other states should be used, such that the representations at each step are learned in context of each other. This comes not only with the advantage that the model has knowledge about all elements of the input sequence at once but also provides the benefit that the computational steps required regarding each state can be executed in parallel.

The calculation of the attention scores is realized with queries, keys and values. Thereby the keys are instances of the sequence each containing a value whereas the query is the specific instance that is processed. The query is compared to all keys by calculating the similarity, yielding the most similar keys with respect to the query. Through applying a softmax function, an attention weight is obtained for each query-key pair. The attention score is then calculated by summing up the values of the keys weighted by their corresponding attention weights.

Self-Attention can be further extended to Multi-Head Attention which consists of multiple heads each possessing an independent attention mechanism focusing on different aspects of the sequence. The results of the different attention mechanisms are combined to obtain the attention score.

Figure 1 shows the full architecture of a transformer model. The decoder (right part) has two Multi-Head Attention blocks. The first one uses self-attention with a masked sequence since during generation the model should only have access to previously generated parts. The second one is dedicated to decoder-encoder attention using queries from decoder states and keys and values from encoder states. The feed forward blocks serve to process the output from the attention blocks and the Add & Norm blocks handle residual connections and layer normalization. Residual connections simply add the input of a block to its output and ease the gradient flow through the network. Layer normalization (Ba et al., 2016) is applied to normalize vector representations which is helpful to improve convergence stability. Since the transformer model does not contain recurrence, it has no knowledge about the order of the sequence. Therefore, Positional Encoding adds embeddings containing information about the positions to the input and output sequence. In most transformer architectures encoder as well as decoder blocks are stacked on top of each other (indicated by $N\times$ in the figure) whereby the output of one block is the input for the following.

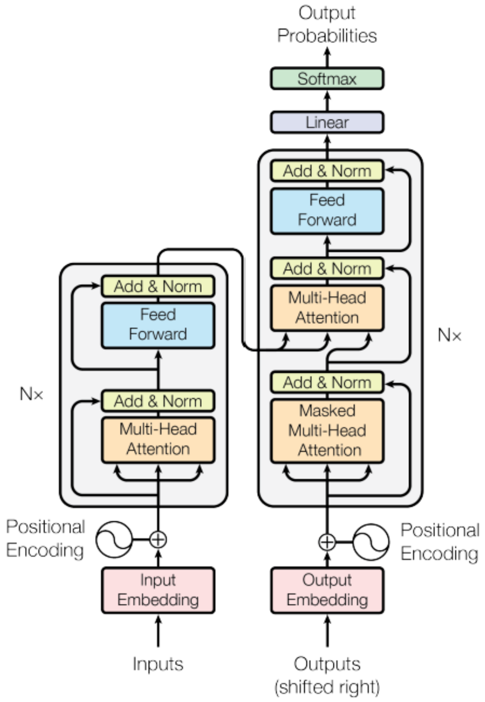


Figure 1: Transformer Model Architecture

Taken from Vaswani et al. (2017)

2.1.5 Generative Adversarial Networks

Generative Adversarial Networks (GANs, Goodfellow et al. (2014)) consist of two convolutional neural networks: one that generates outputs (generator) and one that tries to separate the generated outputs from real ones (discriminator). During training these two networks compete against other which increasingly improves their performance until generated outputs are nearly indistinguishable from real ones. GANs have been applied with great success in image generation but are also used in speech synthesis for the conversion of spectrogram to waveform.

2.1.6 Squeeze and Excitation Blocks

Squeeze and Excitation Blocks (Hu et al., 2018) are building components for neural networks and very helpful for explicitly modeling inter-dependencies between channels. The main idea thereby is that informative features within a given feature map are emphasized while less relevant ones are suppressed. This is achieved by two steps: *squeeze* and *excitation*. Figure 2 shows an overview of the whole pipeline.

In the *squeeze* step the global information of the feature map is captured by applying global average pooling which calculates the average value of each channel. This essentially encodes the importance of each channel in context of the entire feature map.

The goal of the *excitation* step is to model inter-dependencies between channels. This is achieved by a neural network consisting of fully connected layers. The output of these layers is used to generate a channel-wise excitation signal representing the learned importance of each channel. This signal is then applied to the original feature map to modulate channel responses.

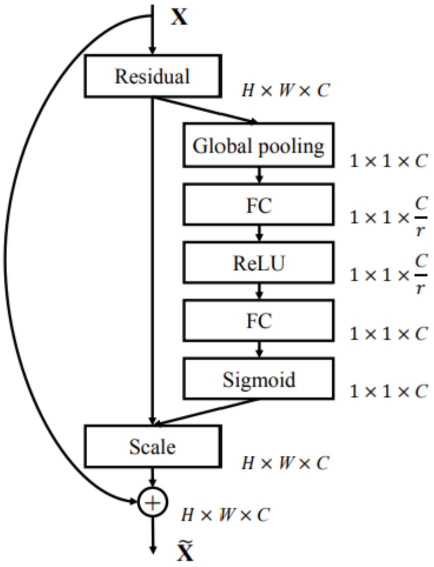


Figure 2: Squeeze and Excitation Block

Taken from Hu et al. (2018)

2.2 Large Language Models

In general language models aim to model the likelihood of word sequences and predict the probabilities of future or missing tokens in a given text. Early approaches such as n-gram models are based on statistical learning methods and try to compute the probability of a word given its previous n context words. These probabilities are based on text corpora from which co-occurrence counts of words are extracted. While such models can be useful for natural language processing tasks like information retrieval due to their efficiency and interpretability, they have several disadvantages. The main drawback is that the computational cost exponentially increases with the number of context words and therefore long-term dependencies cannot be modeled. Furthermore they suffer from a data sparsity problem since even in large text corpora there are a lot of low frequent n-grams. Smoothing strategies such as back-off estimation or Good-Turing smoothing (Chen and Rosenfeld, 2000) have been proposed to tackle this problem.

More advanced language models learn the probabilities of word sequences through neural networks. Thereby the idea of representing words as vectors in a multidimensional space comes more into focus. This concept is based on the distributional hypothesis which states that words occurring in similar contexts tend to have similar meanings (Harris, 1954). One of the first neural language models was proposed in Bengio et al. (2003). It consists of a feed forward neural network used to predict the next token given a sequence of words, whereby a distributional embedding matrix is learned. Also using a feed forward neural network, Mikolov et al. (2013) introduced novel techniques known as Continuous Bag-of-Words (CBOW) and Skip-Gram that improve the models ability to capture semantic relationships between words. While the objective of CBOW is to predict the target word given its surrounding context words, Skip-Gram aims to predict the context words from a given target word. Another approach called Global Vectors for Word Representation (GloVe) was proposed by Pennington et al. (2014) with a focus on capturing global dependencies between words. This is achieved by constructing a co-occurrence matrix from a large corpus and using it during training with the objective of capturing the co-occurrence probabilities between word pairs.

Recurrent neural networks are also used for language modeling (Mikolov et al., 2010) with the advantage of being able to better model the sequential nature of text. Since the vanishing gradient problem hinders the modeling of long-term dependencies, state-of-the art methods make use of the attention mechanism and especially the transformer architecture. These models can be unsupervisedly and efficiently pre-trained on large amounts of text data, e.g. scraped from the internet and thus are called large language models. The most popular models hereby are BERT (Bidirectional Encoder Representations from Transformers, Devlin et al. (2019)) and GPT (Generative Pretrained Transformer, Radford et al. (2018)). BERT is a transformer network trained on the masked language modeling task where the objective is to predict a masked token in a piece of text by using its surrounding context. A second training objective is the prediction of the next sentence where the model is given two sentences and learns to predict if the second sentence follows the first one. GPT is also based on the transformer architecture but differently from BERT its training objective is to predict the next token in a piece of text, so it only uses previous context.

In contrast to models like word2vec and GloVe which learn static word embeddings that are constant regardless of the context in which the words appear, transformer based language models learn contextualized representations. This means that the embedding for a specific word is generated dynamically based on its context via self-attention which can be beneficial for e.g. resolving ambiguities.

Since these transformer based models have shown strong performance in various natural language processing tasks and thus proven their usefulness in understanding and generating human language, numerous model variants have been proposed over the last years. Those models improve the training procedure, increase the number of trainable parameters or specialize on specific domains. Examples for such models are RoBERTa (Robustly Optimized BERT Pretraining Approach, Zhuang et al. (2021)), GPT-3 (Brown et al., 2020), T5 (Raffel et al., 2020), BLOOM (BigScience-Workshop, 2022), LLaMa (Touvron et al., 2023). A noteworthy, recently very successful application of the GPT model is ChatGPT [4] which is adapted on dialogue scenarios and includes reinforcement learning from human feedback (Ziegler et al., 2019) in its training procedure.

2.2.1 Sentence Embeddings

While it is common to extract word embeddings from large language models it is also possible to obtain embeddings that represent the meaning of a whole sentence or piece of text. The most intuitive way of obtaining such a representation is to combine the embeddings of all occurring words. Thereby those can be summed up, averaged or min/max

[4] `https://openai.com/blog/chatgpt`

pooled. Another possibility is to extract the embedding of a token prepended to the sentence commonly referred to as *CLS* (classification). While some research doubts that the *CLS* embedding is entirely appropriate for capturing the semantic representation of a sentence (Tyagi et al., 2020) per se, it is usually used for fine-tuning the model on a classification tasks and thereby captures meaningful information about the sentence. Furthermore there are models specifically trained to learn the meaning of sentences. In this vein Reimers and Gurevych (2019) develop sentence-BERT which modifies BERT by using siamese and triplet networks to derive semantically meaningful sentence embeddings.

2.3 Prompting Methods

In recent years an increasing amount of research has focused on prompting methods and successfully applied them to solve natural language processing tasks (Liu et al., 2022; Radford et al., 2019; Petroni et al., 2019). The prompting paradigm is based on the assumption that the task at hand can be formulated as a masked language modeling or natural language generation problem. The key idea is to manipulate the data by providing a prompt which triggers a pretrained large language model to solve the task. A great benefit of prompting is that it obliterates the necessity to fine-tune the model on the downstream task and therefore also the need for task specific data for training. Instead the data is manipulated to fit the already pretrained model, e.g. Radford et al. (2019) append *"TL;DR"* to the end of a sentence and let a language model generate a summary.

Prompting has also achieved great success in image generation where a prompt is used as a description of the desired image to guide the model (Reed et al., 2016; Rombach et al., 2022; Ramesh et al., 2022).
Thereby providing an appropriate prompt has a large effect on the model's performance. Thus different methods have been developed to choose a proper prompt, in research referred to as prompt engineering. The most intuitive way is to create prompts manually and probe the outcome. While this can be time consuming, another drawback is that the model might be sensitive to the prompt such that even relatively similar inputs can produce high variance in the outcome (Jiang et al., 2020). In order to tackle these problems various methods have been proposed to automatically create prompts which yield a high performance (see Liu et al. (2022) for an overview). However these methods are not readily applicable to text-to-speech synthesis and are therefore not discussed in detail. Instead the primary

goal in this work is to examine if it is possible to guide the generated speaking style and prosodic properties of a text-to-speech system by providing appropriate natural language prompts.

2.4 Speech Prosody

In phonetics the properties of speech are distinguished in segmental and supra-segmental features. While segmental features are dedicated to the pronunciation of specific phonemes, supra-segmental features describe the articulation across larger units of speech, e.g. whole utterances. The most important supra-segmental features are pitch (fundamental frequency), loudness (power in spectrum), voice quality (energy distribution in spectrum) and tempo (phoneme duration and pauses). These features are collectively referred to as speech prosody (Xu, 2011).

Intonation describes the variation of pitch whereas rhythm characterizes the pattern of inserted pauses and intervals between stressed syllables. Prosodic features may transmit emotional states of speakers, the form of the utterance (statement, question, command), the presence of irony or sarcasm and the utilization of contrast and focus.

2.5 Emotion Theories

From an evolutionary point of view emotions focus on observable expressions and have functions that support communication (Darwin, 1872). According to the James-Lange theory emotions arise from the perception of physiological responses to external stimuli (James, 1884; Lange and Kurella, 1885). In contrast the Cannon-Bard theory (Cannon, 1927; Bard, 1928) posits that emotions occur simultaneously with physiological responses and are thus independent but parallel processes.

Besides various models that try to describe emotional states as discrete categories, there exist approaches that represent them continuously along certain dimensions. One of the most popular list of basic emotions was presented by Ekman et al. (1999) through the study of facial expression. This list contains *anger, disgust, fear, joy, sadness* and *surprise*. Plutchik (1980) expands it by including *trust* and *anticipation*. Furthermore he develops a dimensional model by placing the emotions on a wheel with contrasting emotions on opposing sides and inserting more fine-grained distinctions for each basic emotion according to their intensity. Another popular dimensional model is the continuous circumplex model (Russell, 1980) which consists of a two-dimensional vector space with valence on one axis and arousal on the other. Valence describes the degree of pleasantness, ranging from negative to positive while arousal defines the amount of activation, i.e. intensity, and ranges from calm to excited (Bestelmeyer et al., 2017). While this model allows to represent emotional states as points in space, basic emotions can also be mapped into it, e.g. *joy* could be placed somewhere in the region of high valence and high arousal.

A more recent emotion theory by Barrett (2017) emphasizes that emotional states are not fixed categories but rather dynamic and context-dependent experiences, e.g. valence and arousal are perceived and based on context the brain predicts which emotion makes sense.

2.6 Text-to-Speech Synthesis

Text-to-speech synthesis describes the process of generating human like speech given an arbitrary text as input.

In traditional approaches the input text is first processed by a text analysis module which is responsible for pre-processing steps such as word disambiguation, detection of abbreviations and extracting linguistic features like part-of-speech tags and prosodic predictions. The extracted features are passed to the synthesis module which generates speech as a waveform. Popular techniques following this approach are concatenative (Moulines and Charpentier, 1990; Hunt and Black, 1996) and parametric speech synthesis systems (Yoshimura et al., 1999; Tokuda et al., 2000; Black et al., 2007).
However the pipeline of these approaches is rather complex and the linguistic feature extraction can often be expensive. Furthermore the quality of synthesized audios is mostly significantly worse than human speech because it tends to be unstable in prosody and pronunciation, thus sounding unnatural.

In past years the rapid advances in machine learning and in particular deep neural networks have opened the door for end-to-end generative text-to-speech synthesis models. They simplify the traditional pipeline by converting the input text into phonemes and thus removing the need for a text analysis module. In most models the generated speech is first represented as a mel-spectrogram predicted from the input phonemes through a neural network and then transformed into a waveform by a separately trained model known as vocoder.

A spectrogram represents the amplitude of an audio signal as it varies over time and frequencies. It can be obtained by applying short-time Fourier transformation which comprises several fast Fourier transformations on windowed segments of the signal (Sejdić et al., 2009). The x-axis corresponds to the time steps whereas the y-axis indicates the frequency values. The amplitude is represented in a third dimension which is usually represented as a gray scale when plotting the spectrogram as 2-D image. The frequency is mapped to a mel-scale which has been introduced by Stevens et al. (1937). It divides the frequency range into bins where each of them contains frequencies that are perceived as equally by listeners. Thereby the underlying motivation is that humans do not perceive frequencies on a linear scale, e.g. small differences in low frequencies are much better detectable compared to differences in high frequencies.

Models like Tacotron (Wang et al., 2017), Tacotron2 (Shen et al., 2018) and TransformerTTS (Li et al., 2019) generate mel-spectrogram frames autoregressively meaning that each generation step depends on previously generated frames. Although they achieve high scores in terms of perceived naturalness, they typically suffer from slow inference speed and robustness issues such as word skipping and repeating words.

In order to avoid these issues, non-autoregressive models such as FastSpeech2 (Ren et al., 2020), Glow-TTS (Kim et al., 2020) and Grad-TTS (Popov et al., 2021) generate mel-spectrogram frames in parallel while maintaining high speech quality. As a consequence of the non-autoregressive workflow these models require some component to predict the duration of each phoneme and match it to the length of the mel-spectrogram, e.g. in Fastspeech2 this is done with a dedicated duration predictor.

Although non-autoregressive text-to-speech systems are able to produce high quality speech which sometimes is even comparable to the one of human speakers, a challenge that still remains is improving its expressiveness, i.e. by producing appropriate prosody. Fastspeech2 introduces a pitch and energy predictor additional to the duration predictor in order to infer the pitch contour in generated speech. However this only works successfully to a certain degree because the prediction of the model is solely based on the phonemes of the input text and thus is not able to use information about the words or the sentence structure themselves.

An existential amount of recent work investigates ways to condition the model on additionally provided information in order to improve the naturalness and controllability of speech prosody (discussed in section 3).

3 Related Work

3.1 TTS Systems Incorporating Supplementary Information

3.1.1 Prosodic Information

The most obvious way to manipulate speech prosody is to directly providing prosodic information to the TTS model. In this vein several approaches extract and condition the model on explicit prosodic features such as pitch, energy and duration (e.g. Shechtman and Sorin (2019); Raitio et al. (2020)). However during inference it can be complicated to control these features since explicit values have to be provided manually which requires phonetic knowledge.

Chen et al. (2021) tackle this problem by introducing a speech BERT model that is trained to predict prosodic attributes in mel-spectrogram given text and a masked speech segment. During inference their model is able to dynamically construct mel-spectrograms from text with fine-grained prosodic information.

Du and Yu (2022) include a prosody predictor in the TTS pipeline by optimizing an additional loss between prosody embeddings predicted by the model and extracted ones from the training data. They use a Gaussian Mixture Model based density network to represent prosody on phone level which allows for more naturalness and diversity.

Cornille et al. (2022) propose a network that encodes the prosodic information of an entire utterance in a single embedding. During inference they are able to control the prosody by injecting embeddings extracted from a reference utterance.

Furthermore there are numerous approaches which extract representations from speech that encode not only prosodic information but also speaker dependent properties such as timbre and general speaking style, e.g.Wang et al. (2018); Jia et al. (2018); Li et al. (2022). While these models demonstrate great performance in cloning speaker dependent speaking style and prosody, they require reference audios during inference that might not be readily available. Furthermore the desired prosodic realization might not be existent for specific speakers.
The latter issue is tackled by approaches investigating cross-speaker style transfer which aims to generate the speaking style and prosodic properties of one speaker in the voice of another (e.g. Wu et al. (2021); An et al. (2022); Zhang et al. (2023)).

3.1.2 Syntactic Information

There a numerous approaches that incorporate syntactical information into TTS systems with the goal of improving the naturalness of speech and obtaining appropriate prosody. The general motivation behind these approaches is rooted in the idea that syntax provides information about sentence structure and the importance of individual words. These properties have been found to be tightly connected to prosody (Wagner and Watson, 2010; Köhn et al., 2018).

Guo et al. (2019) concatenate the phoneme embeddings in a Tacotron architecture with syntactic features based on parse trees and phrase structures. The experimental results show that their approach improves pronunciation clarity, prosody and naturalness of synthesized speech.

Tyagi et al. (2020) also build their TTS system on Tacotron and select an acoustic embedding, which is used for synthesis, with help of syntactic embeddings obtained from constituency parsing as well as semantic embeddings obtained from BERT (Devlin et al., 2019). Thereby they choose the acoustic embedding of a sentence in the training set for which the syntactic or semantic embedding is most similar to the one of the given sentence.

The work of Liu et al. (2021a) includes syntax aware graph attention in TransformerTTS (Li et al., 2019) whereby syntactically motivated character embeddings are obtained from syntax graphs.

Karlapati et al. (2021) predict prosodic embeddings with help of parse trees and BERT embeddings by learning the prosodic distribution from spectrogram during training and sampling from this distribution during inference. Their model is based on DurIAN (Yu et al., 2020).

Setlur et al. (2021) feed linguistic structure in the form of heterogeneous relation graphs to TTS systems based on Tacotron and TransformerTTS and demonstrate improvements on single speaker and multispeaker datasets.

Zhou et al. (2022b) concatenate word embeddings obtained from BERT with encoder outputs from Tacotron2. Thereby the BERT embeddings are reprocessed by taking dependency information into account which results in more effective semantic representations for text-to-speech enriched with syntactical information.

Ye et al. (2022) propose a TTS model called SyntaSpeech based on PortaSpeech (Ren et al., 2021) which includes syntactic embeddings obtained from dependency parsing by a syntactic graph encoder.

3.1.3 Semantic Information

Following the same motivation as including syntactic information to TTS systems, in the past years there has been a substantial amount of research that incorporates semantic information, mostly in the form of word embeddings obtained from large language models. The underlying assumption thereby is that the semantics of words and phrases is relevant to how they are pronounced. Furthermore it shown that word embeddings can capture some degree of syntactic information (Jawahar et al., 2019; Tenney et al., 2019).

Wang et al. (2015) include word embeddings in RNN synthesis and show that those are able to replace features such as ToBI (Tone and Break Indices) and part-of-speech tags.

Hayashi et al. (2019) were one of the first to incorporate BERT embeddings as input to a neural end-to-end TTS system. They explore the usage of word embeddings as well as sentence embeddings and provide them as additional input to the decoder of Tacotron2 through concatenation. Their experiments demonstrate slightly improved naturalness on a single speaker dataset.

Fang et al. (2019) also include BERT embeddings as input for Tacotron2 with the goal of lowering the reliance on high quality training data. While they do not observe significant quality improvements, they find that BERT embeddings are helpful for faster convergence during training and guiding the decoding process.

Ming et al. (2019) include word embeddings obtained from a neural machine translation task into Tacotron2 and observe improved generalization ability and robustness in addition to quality improvements for out of domain text.

Xiao et al. (2020) investigate the usage of BERT character embeddings in a multi-speaker setting. Thereby phone embeddings and character embeddings are first concatenated and encoded before introducing speaker embeddings. This prevents an equal treatment of the different sources which they claim is beneficial. They apply their model on news, chat and audio-book data and show its effectiveness for improving prosody.

Kenter et al. (2020) follow a similar approach as Wang et al. (2015) and include BERT embeddings in RNN TTS. They find that it is crucial to fine-tune the parameters of BERT while training the model in order to boost results.

Shen et al. (2021) base their work on Fastspeech2 and suggest to include acoustic word embeddings which are jointly trained with the TTS model. They reason that linguistic word embeddings (e.g. obtained from BERT) are not directly relevant to how words are pronounced which is in line with the observation of Kenter et al. (2020) emphasizing the need for fine-tuning or adaptation. Furthermore they add a prosody predictor to their model architecture which can improve naturalness and provides a convenient way to objectively evaluate prosody predictions. Subjective evaluations regarding speech naturalness suggest that their approach is beneficial over TTS systems incorporating plain word embeddings.

Xu et al. (2021) use sentence embeddings of neighboring sentences as additional input to Tacotron2. Besides improved naturalness and expressiveness they observe that it is possible to subtly control the prosody by changing the context sentences.

Cong et al. (2021) use BERT character embeddings to improve conversational speech and achieve more natural prosody while Zhang and Ling (2021) predict prosodic embeddings for each word. These word-level style variations are a combination of embeddings learned from mel-spectrogram during training and semantic information of BERT embeddings. They report improvements over a Tacotron2 baseline on a Mandarin Chinese dataset and a model following the approach of Hayashi et al. (2019) directly using BERT embeddings for synthesis.

3.1.4 TTS with Prompting Capabilities

One of the first approaches allowing prosodic controllability of synthesized speech intuitively with natural language descriptions is style tagging TTS by Kim et al. (2021b). They train a TTS model on a dataset containing manually provided style tags along with the utterances and transcriptions. The model is optimized to predict style representations from these descriptions by employing a loss between embeddings of style tag and reference speech. In order to transfer the style tags into an embedding space they use SBERT (Reimers and Gurevych, 2019), which is a fine-tuned version of BERT optimized for sentence level similarity. Its great generalization capabilities enable the TTS model to process unseen style tags during inference. Sentence embeddings are provided as additional input to the decoder and duration predictor of the model. The bi-modal embedding space learned during training allows to either reference speech or style tag to control prosodic properties of generated speech. Experimental results show that style tagging TTS improves speech quality and expressiveness compared to a Tacotron2 baseline integrating GST (Wang et al., 2018).

Following a similar approach, Shin et al. (2022) propose a bi-modal style encoder that learns the relationship between text style embeddings and speech style embeddings by introducing a loss between them. The text embeddings are also obtained using pre-trained sentence BERT (Reimers and Gurevych, 2019). Differently to Kim et al. (2021b), they additionally introduce a loss between style vectors extracted from reference and predicted spectrogram along with a contrastive loss between style text embedding and embedding extracted from predicted spectrogram. In their experiments they report respectable performance on emotional style control as well as cross-speaker style transfer even for unseen styles for specific speakers.

Wu et al. (2022) build a framework that unsupervisedly learns semantic representations focusing on emotional content from unlabeled text. It combines deep embedded clustering with contrastive learning through data augmentation whereby emotional words in a sentence are replaced by similar ones using an emotion lexicon. In their experiments the framework is pre-trained on a text dataset collected from e-books and is then further used to encode the style of utterances from a TTS dataset consisting of audio-books. A TransformerTTS model is trained on the same dataset with style embeddings as additional conditioning signal. The results suggest that their approach achieves natural expressiveness and emotion transition especially for long paragraphs.

Also focusing on emotional text-to-speech, Mukherjee et al. (2022) propose a text-aware FastSpeech2 system conditioned on embeddings obtained from a BERT model which is fine-tuned to predict emotion categories. They observe that in most emotional speech datasets the emotional content of the text does not match the one of the audio which results in inappropriate BERT embeddings for training. They suggest to circumvent this issue by learning to predict the audio emotion label from the text embedding which can serve as a classifier that detects samples with audio-text emotion disagreement. The results of their experiments show that their model achieves improved emotion accuracy over a baseline that uses explicit emotion labels while maintaining overall speech quality.

PromptTTS (Guo et al., 2022b) represents another TTS model that allows to control the style of synthesized speech with natural language descriptions. It is trained on a dataset containing natural language prompts with style information along with corresponding speech. Their style embeddings are extracted from pre-trained BERT based on the *CLS* token which is fine tuned on the prediction of pre-defined labels that are provided in the dataset and comprise gender, pitch, speaking speed, volume and emotion. Furthermore they integrate the style embeddings to each transformer block in the encoder and decoder. In their experiments they use a generative model to obtain more style prompts with similar semantics and achieve precise style control with high speech quality.

Tu et al. (2022) enrich an emotional speech dataset with contextual information by merging it with an emotional text dataset. Thereby an appropriate context sentence for each utterance is found by matching the emotion labels of both datasets. This approach eliminates the need for manually annotating utterances with suitable style tags or descriptions. The proposed TTS model comprises an encoder-decoder transformer architecture based on OFA (One For All, Wang et al. (2022)) used for spectrogram prediction. The text embeddings of context sentences are concatenated with phoneme embeddings extracted from transcriptions of the utterances before going into the encoder. The text embeddings are initialized with those provided by OFA and fine-tuned during training. The experiments show that the proposed model can generate high-quality and expressive speech based on the given context in both in-domain and out-of-domain scenarios.

InstructTTS (Yang et al., 2023) follows an approach similar to Guo et al. (2022b) by training a text-to-speech system on a dataset containing natural language style prompts. In contrast to PromptTTS they do not fine-tune their sentence embeddings on predefined labels and claim that their system allows more freedom for designing prompts since the style descriptions are not as much constraint in their form. The proposed style encoder module takes embedded style prompt and content as well as encoded audio into account. In order to effectively minimize the distance between style prompt embedding and encoded audio, cross modal representation learning is applied. Furthermore, to prevent the style encoder from capturing speaker or content related information, style-speaker and style-content mutual information are minimized whereby speaker information is provided through an embedding look-up table. In their experiments they perform extensive subjective and objective evaluations supported by ablation studies and show that their model is able to produce high-fidelity and natural speech. At the same time they demonstrate the controllability of the speaking style through textual prompts.

Liu et al. (2023) propose PromptStyle which is another approach in the direction of controlling speech style through natural language descriptions. Their model builds on VITS (Kim et al., 2021a) and incorporates a cross-modal style encoder which learns a shared embedding space of prompt embedding from text and style embedding from speech through cosine similarity loss. The training procedure is divided into two stages whereby in the first stage the model is trained only with style embeddings from speech to learn a useful embedding space for style transfer. In the second stage, semantic embeddings extracted from textual prompts are included with the cross-modal style encoder. These embeddings are obtained from a pre-trained BERT model with an additional adaptation layer. The model additionally receives multi-speaker controllability by feeding speaker embeddings from a look-up table to the decoder. Similar to Shin et al. (2022), during inference the desired style can be controlled either by providing reference audios or textual prompts. The suggested two-stage training approach reduces the required amount of data annotated with style descriptions and experiments on audio-book data demonstrate proper style transfer while preserving high speech quality and speaker similarity.

TorToise (Betker, 2023) is a text-to-speech system with great voice cloning capabilities that adopts recent methods from image generation and is trained on large amounts of data. In preliminary research it has been found that the model can be prompted with simple style descriptions such as *”I am really sad”* in order to influence the speech prosody of the utterance that is synthesized. While the system has no specific design that handles the controllability during training, it can be hypothesized that the connection between semantics and prosody is learned through the vast amount of data.

3.2 Vocoders

The spectrograms that are predicted by TTS systems have to be converted into waveform in order to obtain an audio signal. This can be achieved by a so-called vocoder which is trained separately from the TTS system. State-of-the art vocoders are based on generative adversarial networks (GAN) which consist of a generator and a discriminator that are trained adversarially. The generator up-samples the mel-spectrogram through transposed convolutions until the temporal resolution of raw waveform is reached. Then the discriminator decides if the audio sample is real, i.e. obtained from training data or fake, i.e. produced by the generator.

Kumar et al. (2019) introduce MelGAN which applies techniques that allow reliable generation of high quality coherent waveforms from mel-spectrograms. The authors suggest using a multi-scale discriminator operating on different scales of the input waveform combined with spectral normalization which helps to stabilize training. Furthermore, they propose a feature matching loss, which encourages the generator to produce waveforms that match the statistics of intermediate features extracted by the discriminator. Finally, they use a history buffer which stores generated samples and prevents the discriminator from over-fitting to the generator.

HiFiGAN (Kong et al., 2020) achieves generation of high-fidelity speech by additionally modeling periodic patterns of audio by a separate discriminator.

Avocodo (Bak et al., 2022) also adopts two discriminators (multi-band and sub-band) which reduces the emergence of artifacts while maintaining fast inference speed and high quality.

BigVGAN (Lee et al., 2022) is a large-scale GAN that further improves audio quality by establishing periodic activation function and anti-aliased representation in the generator.

3.3 IMS Toucan Speech Synthesis Toolkit

The IMS Toucan Speech Synthesis Toolkit was first introduced in Lux et al. (2021) and presents open-source code for non-autoregressive speech synthesis adopting architectural designs from state-of-the art models such as TransformerTTS (Li et al., 2019), FastSpeech2 (Ren et al., 2020) and FastPitch (Lancucki, 2021). In Lux et al. (2022) it is extended for low-resource multilingual and zero-shot multi-speaker TTS using language agnostic meta-learning (Lux and Vu, 2022). Koch et al. (2022) investigate the prosody cloning capabilities of the toolkit with regard to controllable poetry reading for literary studies. Lux et al. (2023b) further demonstrate that it is possible with IMS Toucan to clone the voice of a speaker as well as the prosody of a spoken reference independently. The most recent contributions up to the submission date of this thesis have been made to the toolkit for the Blizzard Challenge 2023 (Lux et al., 2023a). The baseline and proposed approach as described in section 4 are heavily based on this version of the toolkit. Overall, while there might be TTS systems producing more natural speech, the IMS Toucan system offers benefits with respect to speed, data efficiency, robustness and controllability.

4 Methods

4.1 Baseline

The baseline model architecture is build on the version of the IMS Toucan Speech Synthesis Toolkit submitted to the Blizzard Challenge 2023 (Lux et al., 2023a). While its general architecture is based on FastSpeech2, it also adopts components from PortaSpeech (Ren et al., 2021) and FastPitch (Lancucki, 2021). The system produces speech in multiple steps which comprise processing the text, predicting spectrogram frames and transforming spectrogram into waveform. An overview of the whole pipeline can be seen in figure 3. The corresponding implementation and models can be accessed on GitHub [5].

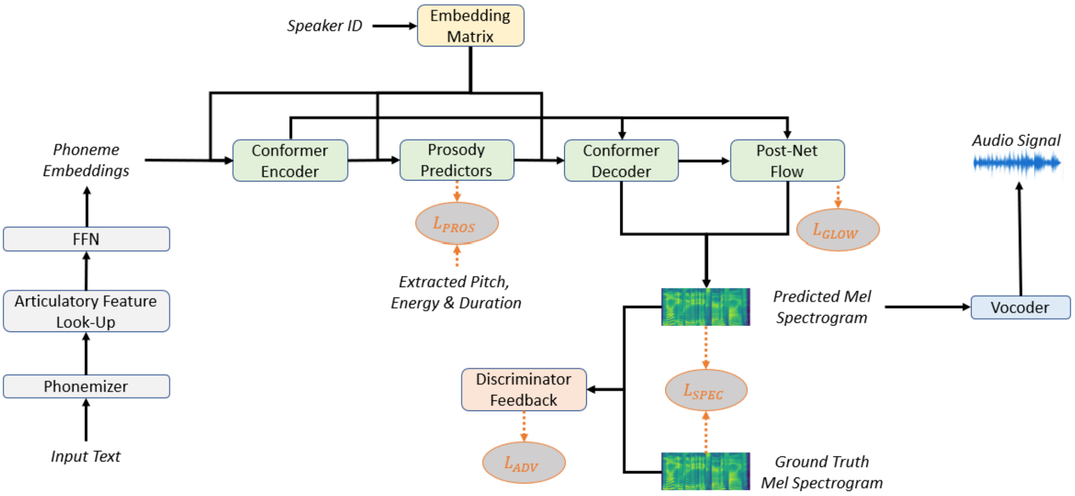


Figure 3: Baseline Architecture

[5] `https://github.com/Thommy96/IMS-Toucan`

4.1.1 Text-to-Phoneme Conversion

The input text is processed and converted into a sequence of phonemes in IPA notation (International-Phonetic-Association, 1999) using an open source phonemizer [6] with espeak-ng [7] as back-end. This step also includes rudimentary text cleaning and the expansion of abbreviations. Each phoneme is then further transformed into a vector containing information about the configurations of the human vocal tract in a one-hot encoding style as introduced in Lux and Vu (2022). The vectors are obtained using a lookup-table indicating which features are present for each phoneme. The feature set consists of 62 entries derived from Staib et al. (2020) and includes phonological features describing IPA phonemes, e.g. voicing, place of articulation or the positioning of the tongue, as well as supplementary dimensions corresponding to non-segmental markers such as lengthening, shortening and lexical stress. These dimensions were introduced to the system in Lux et al. (2022) in order to account for tonal languages and are modified by previous or following units. The main purpose of having phoneme representations described by phonological features is to be able to handle unseen phonemes and to learn phoneme representations that are applicable across multiple languages. While this facilitates the architecture for multi-language systems, they found that their approach is also beneficial for single-language cases due to knowledge sharing between phonemes making the system converge much earlier.

4.1.2 Mel-Spectrogram Extraction

Ground-truth mel-spectrograms are extracted from waveform after applying some signal processing such as loudness normalization using the pyloudnorm tool (Steinmetz and Reiss, 2021) and cutting silence from the beginning and end of the audio. The amplitude spectrograms are obtained from waveform by performing a short-time-Fourier-transform (STFT) with a window size of 1024, a hop length of 256 and a Hann window. They are further transformed into mel-spectrograms with 80 frequency bins using the Librosa toolkit (McFee et al., 2015) and finally a logarithm with base 10 is applied which results in more convenient value ranges.

[6] https://github.com/bootphon/phonemizer

[7] https://github.com/espeak-ng/espeak-ng

4.1.3 Spectrogram-Phoneme Alignments

Since the underlying FastSpeech2 system is trained in parallel, it is crucial to predict the duration for each phoneme. Furthermore this prediction is necessary since pitch and energy values are averaged over phoneme durations. Inspired by Pérez-González-de Martos et al. (2021) these durations are obtained by first training an auto-encoder framework prior to the actual TTS system from which alignments of phonemes to spectrogram frames can be extracted, i.e. it gives information about which and how many spectrogram frames correspond to each phoneme. An overview of the framework's architecture can be seen in figure 4.

It consists of two interconnected modules which are trained in an end-to-end manner. The first module is a simple recurrent neural network based speech recognition system with a Connectionist Temporal Classification (CTC) objective (Graves et al., 2006). Formally, CTC is a way to produce alignments between an input sequence $X = (x_1, x_2, ..., x_T)$ and an output sequence $Y = (y_1, y_2, ..., y_U)$. Thereby the target sequence cannot be longer than the input sequence, i.e. $U \leq T$. The CTC objective can be formulated as in equation 3, where $p(Y|X)$ describes the conditional probability for the target sequence Y given the input sequence X. It is computed by estimating the probability for each single alignment a_t at time step t and marginalizing over the set of all valid alignments $A \in \mathcal{A}_{X,Y}$.

$$p(Y|X) = \sum_{A \in \mathcal{A}_{X,Y}} \prod_{t=1}^{T} p_t(a_1|X) \tag{3}$$

This means that the network provides a distribution of posterior probabilities $p_t(a|X)$ over the elements of the output sequence at each time step t, also called posteriorgram.
In the speech-to-text (STT) module the input spectrogram frames are first processed by a stack of five 1-D convolutional layers with batch normalization and ReLU activation functions. After that they are fed into a single-layer bi-directional LSTM and finally a linear projection layer with softmax activations is used to output the phoneme probabilities, i.e. the posteriorgram.

This is then further passed to the second module which is a simple TTS system bearing the task of reconstructing the original spectrogram frames. It incorporates a two-layer bi-directional LSTM succeeded by a linear projection layer mapping to the spectrogram dimension. Differently to Pérez-González-de Martos et al. (2021) which use mean absolute error (MAE), in this work the sum of L1 loss (absolute error) and L2 (squared error) is calculated between predicted and ground-truth spectrogram frames. The backpropagation of this loss through the whole framework is helpful for achieving more accurate alignments.

$L1$ loss (equation 4) is calculated as the sum of all absolute differences between predicted and ground truth values while $L2$ loss computes the sum of all squared differences between them (equation 5).

$$L1 = \sum_{i=1}^{n} |y_{true_i} - y_{predicted_i}| \tag{4}$$

$$L2 = \sum_{i=1}^{n} (y_{true_i} - y_{predicted_i})^2 \tag{5}$$

Given the output of the STT module, i.e. the posteriorgram, the most likely alignment can be extracted using Monotonic Alignment Search (MAS) (Kim et al., 2020). This in turn serves then to calculate phoneme durations, i.e. by counting the number of spectrogram frames aligned with each phoneme. While Pérez-González-de Martos et al. (2021) use the Dijkstra algorithm for this purpose, Lux et al. (2023a) found that MAS is beneficial because it is constrained not to skip over segments.

The spectrogram-phoneme alignment framework was introduced to the IMS Toucan system in Lux et al. (2023b) and other work has verified its accuracy and usefulness for TTS (Koch et al., 2022; Meyer et al., 2023; Zalkow et al., 2023).

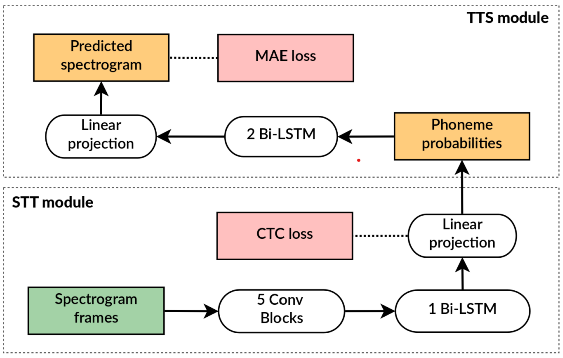


Figure 4: Aligner Framework Architecture

Taken from Pérez-González-de Martos et al. (2021)

4.1.4 Pitch and Energy Extraction

As proposed in Lancucki (2021) pitch and energy values are averaged for each phone according to its duration which results in controllability at phone-level. Moreover pitch values for unvoiced phonemes along with pitch and energy values for symbols corresponding to silence, e.g. pause or punctuation markers, are manually set to zero as to reduce noise. Since the values should be speaker independent, pitch an energy levels are normalized by the mean per utterance while excluding zero values.

4.1.5 Spectrogram Generation

The spectrogram generation framework builds the core of the TTS system and is based on the architecture of FastSpeech 2 (Ren et al., 2020) further augmented with techniques partly inspired by FastPitch (Lancucki, 2021) and PortaSpeech (Ren et al., 2021).

First, the articulatory feature vectors describing the phonemes are passed through two linear layers with Tanh activations in between such that they are embedded for the TTS case and mapped to the hidden dimension size of the encoder-decoder network which is based on the conformer architecture (Gulati et al., 2020).

The conformer combines convolutional neural networks and transformer which enables it to model local as well as global dependencies of the input sequence and has been found to significantly outperform both transformer and convolutional based models (Gulati et al., 2020; Guo et al., 2021). After applying positional encoding to the phoneme embeddings, they are passed through several conformer encoder blocks which are inspired by the Macaron-Net (Lu et al., 2019) with two 2-layer 1-D convolution modules enclosing the multi-head self-attention module and the actual convolution module as can be seen in figure 5. Gulati et al. (2020) propose to use feed-forward modules at the beginning and end of each block, however, based upon the approach in FastSpeech (Ren et al., 2019) they were replaced by convolutions which has shown to improve the quality of generated speech.

The multi-head self-attention module is derived from Dai et al. (2019) and incorporates relative positional encoding which generalizes better over variable input size and thus makes it more robust to different utterance lengths.

The architecture of the convolution module can be seen in figure 6. It consists of a pointwise 1-D convolution with an expansion factor of 2 on the number of channels combined with a gated linear unit (GLU, Dauphin et al. (2017)). This is followed by a 1-D depthwise convolution with batch normalization and swish activation. The final layer of this module is again a 1-D pointwise convolution.

Each module in the conformer block is prepended by layer normalization and after each one residual connections are added whereby the output of the macaron convolution modules is halved.

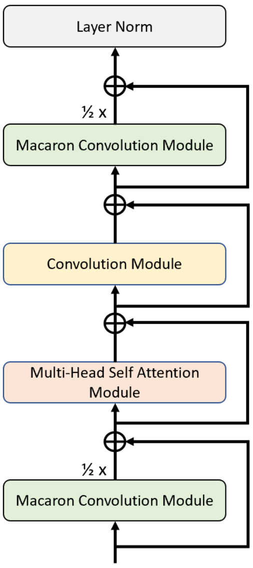

Layer Norm
½ x
Macaron Convolution Module
Convolution Module
Multi-Head Self Attention Module
½ x
Macaron Convolution Module

Figure 5: Conformer Block

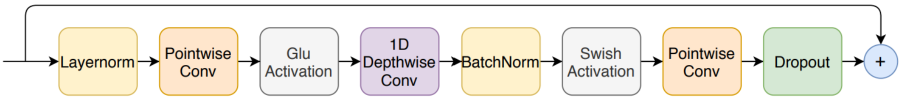


Figure 6: Convolution Module of a Conformer Block

Taken from Gulati et al. (2020)

After applying layer normalization, the output of the encoder is used for the prediction of duration, pitch and energy which corresponds to the variance adaptor in FastSpeech2.
The duration predictor estimates how many spectrogram frames belong to each phoneme and thereby uses the extracted durations from phoneme-spectrogram alignment as ground-truth during training. It consists of two 1-D convolutional layers, each followed by ReLu activation and layer normalization and is optimized with L2 loss. The durations and the loss are predicted in logarithmic domain which makes the value range easier to handle. The pitch and energy predictors have the same architecture as the duration predictor and are also optimized with L2 loss. Ground-truth values for pitch and energy are obtained as described in section 4.1.4.
During inference the predicted pitch and duration values are modified with linguistic knowledge whereby pitch values for unvoiced phonemes and duration values for word boundaries are set to zero. During training the ground-truth values for all three prosodic properties are used in order to train the predictors with teacher forcing. After embedding pitch and energy values through a 1-D convolution layer they are added to the output of the encoder. Since these enriched hidden representations are still on phoneme-level, they have to be expanded to spectrogram frame-level. This is done by a length regulator as introduced in Ren et al. (2019) which repeats each hidden representation based on the corresponding duration values.

The upsampled sequence of embeddings is further passed to the decoder which has the task of generating spectrogram frames for each embedding and follows the same conformer architecture as the encoder. The output of the decoder is passed through a linear layer which maps to the spectrogram dimension.

The predicted spectrogram is further improved by passing it through a 5-layer convolutional postnet as proposed in Shen et al. (2018) with residual connection. The postnet helps to compensate the detailed structure of the spectrogram by refining the predicted mel-filter-bank. In addition a glow-based postnet using normalizing flows is introduced inspired by the work of Ren et al. (2021). Normalizing flows have also been used in Kim et al. (2020) and Miao et al. (2020) with great success in generating high-quality spectrograms. They are able to model fine-grained details and therefore vanquish the over-smoothing problem which often causes blurry outputs. The post-flow usually requires warm-up and is only included after several training steps.

Finally, the TTS system is conditioned on a discriminator feedback. The discriminator is an adversarial network trained along with the TTS system and has the task to distinguish between real and generated spectrograms. It is optimized with discriminator and generator loss obtained by the calculation of mean squared error (MSE).

The overall training objective of the TTS system can be described by the loss function in equation 6 which is the sum of the loss functions of the respective components.

$$\begin{aligned} \textit{TTS-Loss} = {} & \textit{Spectrogram-Reconstruction-Loss} \\ & + \textit{Prosody-Predictor-Loss} \\ & + \textit{Glow-Loss} \\ & + \textit{Adversarial-Loss} \end{aligned} \tag{6}$$

The *Spectrogram-Reconstruction-Loss* is a sum of *L1-Losses* between ground-truth spectrograms and generated spectrograms before and after the post-flow. The *Prosody-Predictor-Loss* is a sum of the losses on which the duration, pitch and energy predictors are optimized. In order to reduce the L1 and L2 losses to a single value, they are averaged by summing up the relevant values respectively and dividing them by the number spectrogram frames or, in the case of the duration predictor, the length of the phoneme sequence. This calculation corresponds to mean absolute error (MAE) and mean squared error (MSE). The *Glow-Loss* describes the loss of the glow-based post-flow and the *Adversarial-Loss* corresponds to the sum of the discriminator and generator loss required to train the adversarial network.

4.1.6 Speaker Embeddings

In order to make the system capable of accurately producing speech in the voice of multiple speakers, a 2-D speaker embedding matrix is instantiated with one dimension corresponding to the embedding size and the other to the number of speakers. During training the embedding vector of the current speaker is selected and updated accordingly with the whole system's backpropagation. Whilst inference the desired speaker timbre can be selected by simply specifying the appropriate index of the learned speaker embedding matrix. While such an approach has been widely used in TTS systems due to its simplicity and robustness (Ping et al., 2017; Chen et al., 2020), its main drawback is that only speakers seen during training can be used reliably during inference meaning that it cannot generalize to unseen speakers. This is caused by the fact that the speaker variety in the training data is usually not high enough. There are several approaches which overcome this weakness by separately training a dedicated model on audio data containing a large amount of speakers (Snyder et al., 2018; Desplanques et al., 2020). Such models have great zero-shot capabilities and can thus be used to extract speaker embeddings for arbitrary audio data. These can then be used for the training of the TTS system while staying fixed, which enables the model to generate speaker timbres unseen during training by providing an appropriate reference audio from which the desired speaker embedding can be extracted and fed into the model during inference. However, in this work the TTS system relies on a speaker embedding lookup-table learned during training with the main reason for this design choice being that speaker embeddings extracted from pretrained models not only embrace speaker timbre but also prosodic properties such as speaking style which interferes with the research focus of controlling prosody solely by conditioning the model on sentence embeddings. This also justifies the omission of Global Style Token (GST) embeddings (Wang et al., 2018) as proposed in Lux et al. (2023a). Since the speaker embeddings are trained on a variety of speaking styles for each speaker, they cannot capture specific prosodic properties. Instead they are forced to capture information regarding the speaker's timbre.

The speaker embeddings are induced into the architecture of the TTS system at several points after passing them through a linear layer which maps them to the hidden dimension of the model. They are used as conditioning signal for the encoder and decoder as well as for the duration, pitch and energy predictors. For the encoder and decoder a speaker embedding is combined with the input hidden representation by repeatedly concatenating it with each embedding vector and applying a linear projection layer to restore the hidden dimension. For the encoder it is additionally added to the output hidden representation in the same way. In the duration, pitch and energy predictors the speaker embeddings are integrated after each convolutional layer through conditional layer normalization replacing the originally used layer normalization. This has been proposed in Lux et al. (2023a) to increase the responsiveness of the predictors to speaker embeddings.

4.1.7 Waveform Generation

During inference, generated spectrograms are converted to waveform using a setup consisting of a selection of the following neural vocoders: MelGAN (Kumar et al., 2019), HiFiGAN (Kong et al., 2020), Avocodo (Bak et al., 2022) and BigVGAN (Lee et al., 2022).

Inspired by Liu et al. (2021b) a sampling rate of 16kHz is used for spectrograms which is more suitable for spectrogram generation. The vocoder super-resolution is performed by using up-sampling scales mapping to 24kHz waveform signals.

4.2 Proposed Method

4.2.1 Extensions of the Baseline

The TTS system proposed in this thesis follows the architecture of the baseline but is extended to additionally condition the model on information about the content of provided textual prompts. An overview of the proposed approach is presented in figure 7. The implementation along with trained models can be accessed on GitHub [8].

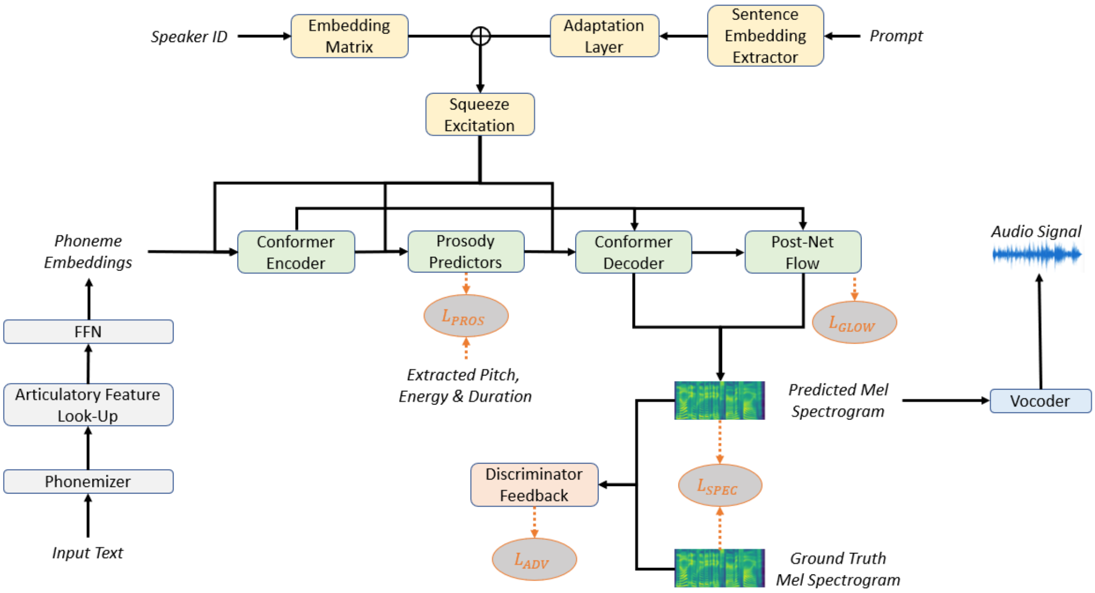


Figure 7: Proposed Architecture

[8] https://github.com/Thommy96/IMS-Toucan

During training, the textual prompt is fed into a sentence embedding extractor based on a large language model which transforms it into a representation capturing the content of the whole input in a single vector. The sentence embeddings are further passed through a linear layer in order to adapt them to the text-to-speech case. Then they are concatenated with the speaker embeddings and projected to the hidden dimension of the system by two linear layers with Tanh activation in between. This results in a representation that contains information about the speaker identity and semantics of the prompt. In the system's pipeline this joint representation is integrated similarly to the speaker embeddings in the baseline.

Furthermore in preliminary tests it has been found that inserting a squeeze and excitation block after concatenating sentence and speaker embeddings increases the controllability of both speaker identity and prosody. It seems that this additional component is beneficial to learn a more meaningful hidden representation. The squeeze and excitation block is able to model inter-dependencies between features of both embeddings and thus helps to disentangle speaker and semantic information. This prevents capturing only one of the both sources when the dimensionality is reduced after concatenation.

By conditioning the speech synthesis model on sentence embeddings of textual prompts, the assumption is that the additional information is useful to predict a fitting prosody. Thereby it has to be assumed that there exists a high correlation between prompts and speech prosody in the training data. This aspect has been found to be crucial since the model has to be able to learn a relation between prompt embeddings and prosodic parameters.

During inference the prompt which is provided to the sentence embedding extractor can be chosen arbitrarily since the large language model is expected to have generalization capabilities. The model should then produce speech with suitable prosody with respect to the prompt as learned during training. This could also be describes as a style transfer from the prompt to the generated utterance whereby no reference audio is needed.

4.2.2 Unsuccessful Designs

While the proposed system uses a speaker embedding look-up table, it has also been tried to use pre-trained speaker embeddings adopted from speaker verification such as Snyder et al. (2018) [9] or Desplanques et al. (2020) [10] instead. This would equip the model with speaker generalization capabilities since these embeddings are trained on a large number of speakers. The core advantage would be that during inference, speech in the voice of unseen speakers could be produced. However because the speaker embedding is extracted from a reference audio, it also captures speaking style and prosodic information of that reference to some extent. Therefore, it is challenging to prevent it from influencing the prediction of prosodic parameters. In this thesis it has been tried to disentangle speaker timbre from prosody by passing the speaker embedding through a tight bottleneck as in Zhang et al. (2023). However produced speech showed a high degradation in speaker similarity. Future work could investigate further approaches tackling this problem such as An et al. (2022). However this has not been implemented here due to time constraints and the research focus on prompts.

Furthermore preliminary experiments included a system additionally integrating word embeddings extracted from the input text. Contextual word embeddings contain information about the semantics of a word in context of the whole sentence. Therefore, they might also capture the importance of specific words and provide guidance regarding how those should be pronounced. Word embeddings were obtained from a standard BERT model by combining the last 4 hidden layers for each sub-token and averaging them over units that form words. In the TTS system each embedding was concatenated with its corresponding phonemes according to the word boundaries provided by the feature vectors. In earlier training stages word embeddings have indeed found to be useful to prevent mispronunciations. However with longer training time there was no significant improvement anymore and therefore this design has been omitted for further experiments.

Apart from that, different integration methods for sentence embeddings have been tried such as providing them independently from the speaker embeddings at different points in the architecture, e.g. only before or after the encoder. It has also been explored to include an additional contrastive loss between sentence embeddings and spectrograms with the idea that similar prompts should result in similar spectrograms. However the much simpler

[9] https://huggingface.co/speechbrain/spkrec-xvect-voxceleb

[10] https://huggingface.co/speechbrain/spkrec-ecapa-voxceleb

approach of concatenating sentence embeddings with speaker embeddings has been found to produce superior results.

Finally it has been attempted to extend the proposed method to a multi-language system. This includes language agnostic meta learning as introduced and applied in Lux and Vu (2022); Lux et al. (2022). While this is in theory well applicable to the proposed system, it was not possible to find suitable expressive speech datasets. Moreover it was not feasible to fine-tune large multi-language models such as *LASER* (Artetxe and Schwenk, 2019) or *LEALLA* (Mao and Nakagawa, 2023) on e.g. emotion prediction in order to make them useful for extracting prompt embeddings. However this could be an interesting direction for future research.

5 Experiments

5.1 Experimental Setup

The goal of the experiments is to test the following hypotheses:

1. Conditioning a speech synthesis model on sentence embeddings of textual prompts improves the quality of generated speech in terms of diverse prosody and naturalness.
2. The prosodic parameters of speech and the conveyed emotional state are controllable by providing suitable prompts during inference.
3. The emotional content of the prompts is correlated to the prosody and conveyed emotion of synthesized speech.
4. Semantically similar prompts result in similar prosody and conveyed emotion.
5. The controllability of speaker identities is not affected by the integration of sentence embeddings.

Hereby it is important to note that the first hypothesis is not the main focus of this thesis since the improvement in speech quality might only be very subtle and therefore hard to measure reliably. The primary goal of this thesis is to investigate hypotheses 2-4, i.e. the intuitive controllability of prosodic parameters of generated speech by providing natural language prompts. The experiments contribute to related work exploring this task while introducing novel training and evaluation strategies within an open-source toolkit, making future research more accessible.

The effect of provided sentence embeddings on the prosodic properties of generated speech is investigated by assessing emotional states. These are one of the most obvious aspects that are expressed by varying prosodic features (Leentjens et al., 1998; Sauter et al., 2010; Pell and Kotz, 2011). For example it has been found that emotions such as fear and anger are related to higher frequencies than sadness (Bachorowski, 1999). Furthermore there exist several emotional speech corpora which can be used for training and testing. They provide a high variance in prosodic properties and are labeled with emotion categories. However, an issue that arises when using such datasets is that the utterances are usually repeated over all emotion categories which means that there is no correlation between the meaning of the sentences and their prosodic parameters when spoken. This correlation is crucial for training the TTS system since the model has to learn that sentences with similar meaning should have similar prosody. In order to circumvent that problem, an auxiliary emotional text dataset is used for the extraction of sentence embeddings, inspired by the approach in Tu et al. (2022). The emotional text dataset has to be labeled with similar emotion categories as the speech dataset such that for each utterance a sentence from the same category can be selected as prompt.

In all experiments the proposed model conditioned on sentence embeddings is compared to the baseline model. Given that the integration of sentence embeddings is the only difference between the two models, its effect can be measured reliably by conducting subjective and objective evaluations which helps to answer the above hypotheses. Audio samples of generated utterances can be accessed online [11] along with an interactive demo on HuggingFace [12].

5.2 Datasets

5.2.1 Speech Datasets

Popular datasets used to train and evaluate text-to-speech synthesis models are LJSpeech (Ito and Johnson, 2017) and LibriTTS (Zen et al., 2019).
LJSpeech is a publicly available [13] English single-speaker dataset consisting of 13,100 short audio clips from a collection of non-fiction books read by a female speaker. Each audio is accompanied with a transcription and the total length of the dataset is approximately 24 hours. This dataset is convenient to train a single speaker text-to-speech system due to its high quality, however, it doesn't provide much prosodic variation.
LibriTTS is a substantially larger corpus of approximately 585 hours of English speech from 2,456 speakers over a large variety of topics accompanied with transcriptions. It is derived from the LibriSpeech corpus (Panayotov et al., 2015) by optimizing it for text-to-speech related work and is open source [14]. Recently Koizumi et al. (2023a) released an enhanced version of this corpus called LibriTTS-R. By applying speech restoration

[11] https://thommy96.github.io/toucanpromptingdemo/
[12] https://huggingface.co/spaces/Thommy96/promptingtoucan
[13] https://keithito.com/LJ-Speech-Dataset/
[14] https://www.openslr.org/60/

with the *Miipher* model (Koizumi et al., 2023b), the sound quality of the speech samples is significantly improved making it more suitable for training a high-quality TTS system. LibriTTS-R is also publicly available [15].

The training of the proposed TTS systems relies on three emotional speech datasets: ESD (Zhou et al., 2021; 2022a), RAVDESS (Livingstone and Russo, 2018) and TESS (Pichora-Fuller and Dupuis, 2020).

ESD (Emotional Speech Database) [16] consists of 10 native English and 10 native Mandarin speakers each uttering the same sentences across five emotion categories. For the experiments in this thesis only the English utterances of five male and five female speakers are used. Thereby the emotion categories are *anger*, *disgust*, *joy*, *neutral*, *sadness* and *surprise*. Each category comprises 350 utterances resulting in 1,750 utterances per speaker, 3,500 per category and 17,500 in total. The total duration of the dataset is approximately 13 hours.

RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song) [17] consists of speech and song data of 24 professional actors whereby the number of male and female actors is balanced. For the purposes of TTS, only the speech data is used which includes utterances across eight emotion categories: *anger*, *disgust*, *joy*, *neutral*, *sadness*, *surprise*, *fear* and *calmness*. Since the training and evaluation is based on Ekman's basic emotions, utterances from the *calmness* category are discarded. Furthermore the dataset provides multi-modality in the form of facial expressions which are also not used. Each speaker produces 8 utterances per emotion category, except for *neutral* where there are only 4. This results in 60 utterances per speaker and 1,440 in total.

TESS (Toronto Emotional Speech Set) [18] consists of utterances from two female speakers each reading sentences in the pattern *"Say the word [X]"* where *[X]* is varied across 200 target words. All sentences were read with respect to seven emotion categories which are *anger*, *disgust*, *joy*, *neutral*, *sadness*, *surprise* and *fear*. Hence the dataset comprises 2800 utterances in total.

An overview of the properties of all used speech datasets can be found in tables 1 and 2.

[15] `http://www.openslr.org/141/`

[16] `https://github.com/HLTSingapore/Emotional-Speech-Data`

[17] `https://zenodo.org/record/1188976`

[18] `https://tspace.library.utoronto.ca/handle/1807/24487`

	LJSpeech	**LibriTTS-R**
# Speakers	1	2,456
Duration in Hours	24	585

Table 1: Speech Datasets

	ESD	**RAVDESS**	**TESS**
# Speakers	10	24	2
# Emotion Categories	5	8	7
# Utterances per Emotion	350	8	200
# Utterances in Total	17,500	1,440	2800

Table 2: Emotional Speech Datasets

5.2.2 Emotional Text Datasets

For training the TTS system conditioned on sentence embeddings it is necessary to provide prompts which are labeled with the corresponding emotion categories. For this purpose sentences from reviews of a subset of the Yelp open dataset (Asghar, 2016) [19] are used. The subset [20] contains 700,000 reviews about businesses in the USA and Canada given by users on Yelp [21]. Thereby the training split of the dataset is used which contains 650,000 instances. Reviews are considered as appropriate for conditioning the TTS system since they contain highly emotional text in common language from various different authors. While a lot of sentences can be extracted from the reviews due to the large size of the dataset, they are not labeled with emotion categories per se. Therefore, an auxiliary emotion classification model (Hartmann, 2022) is used to classify reviews. This is the same model which is used for the extraction of sentence embeddings and is described in more detail in section 5.3.

For each of the seven emotion categories (*anger*, *disgust*, *joy*, *neutral*, *sadness*, *surprise*, *fear*), 10,000 sentences are extracted randomly from classified instances which received a

[19] https://www.yelp.com/dataset

[20] https://huggingface.co/datasets/yelp_review_full

[21] https://www.yelp.com

probability score higher than 0.8 from the model. This results in 70,000 prompts which are used for training the TTS system after extracting their sentence embeddings. Thereby using the same model for classification and extraction is reasonable since the model performs robustly across different datasets (Hartmann, 2022) and only samples which are classified with high confidence are selected.

For the extraction of sentences used during evaluation, two other emotional text datasets are used, both labeled with emotion categories by human annotators.
DailyDialog (Li et al., 2017) [22] contains conversations comprising 13,118 sentences which are crawled from websites where English learners can practice dialog in daily life. Therefore, it contains common language text about diverse topics which at the same time is more formal than e.g. text extracted from Twitter.
The second dataset used for evaluation is Tales-Emotion (Alm et al., 2005; Alm and Sproat, 2005) [23]. It consists of 185 fairy-tales by Potter, Andersen and Grimm with a total of 15,292 annotated sentences.
From both datasets, 50 sentences for each of the five emotion categories *anger*, *joy*, *neutral*, *sadness* and *surprise* are selected. Furthermore sentences that didn't result in a high probability score for the annotated emotion category when applying the emotion classification model were discarded in order to obtain cleaner data.
The conversational nature of the sentences in DailyDialog and the story telling style in Tales-Emotion might demonstrate the usefulness of the TTS model for both dialog systems and reading stories.

5.3 Sentence Embedding Extraction

For the purpose of extracting embeddings from given prompts during training and inference, a large language model based on the DistilRoBERTa architecture [24] is used which is fine-tuned on the task of emotion classification (Hartmann, 2022) [25]. The distilled RoBERTa model follows the same training procedure as DistilBERT (Sanh et al., 2019) which uses the BERT model as teacher for knowledge distillation (Hinton et al., 2015). Thereby two additional training objectives are established on top of the masked language modeling

[22] https://huggingface.co/datasets/daily_dialog

[23] http://people.rc.rit.edu/~coagla/affectdata/index.html

[24] https://huggingface.co/distilroberta-base

[25] https://huggingface.co/j-hartmann/emotion-english-distilroberta-base

task. The first one is a distillation loss which makes sure that DistilBERT returns the same probabilities as the BERT model. The second one is a cosine embedding loss that has the purpose of making DistilBERT generate hidden states similar to BERT. The key advantage of distilled models is that they have less parameters than the originals and thus are faster for inference and downstream tasks while at the same time maintaining comparable performance. DistilRoBERTa was pre-trained on the OpenWebTextCorpus (Gokaslan and Cohen, 2019) and fine-tuned on a collection of emotional text datasets (see Hartmann (2022) for an overview).
The classification is based on the following emotion categories: *anger*, *disgust*, *joy*, *neutral*, *sadness*, *surprise* and *fear*. The authors report an evaluation accuracy of 66% compared to a random-chance baseline of 14%.

Sentence embeddings are extracted from the *[CLS]* hidden representation comprising 756 dimensions. The *[CLS]* state is used for emotion classification during fine-tuning and is therefore expected to capture relevant information about the emotional content of a sentence. Alternatively the provided code offers the possibility of extracting embeddings averaged over the last hidden state or the second to last hidden state of each token, however through internal testing it has been found that the *[CLS]* representation works best.

5.4 Training Procedure

Before training the models, the applied datasets are cleaned by calculating the TTS-Loss for each individual sample and heuristically removing samples which have a high loss until there are no more outliers. This is important since the training data might contain samples with a lot of background noise or mispronunciations.

The training procedure of both the baseline and the proposed model is divided into a pre-training and a fine-tuning stage whereby the difference is that the pre-training stage includes more training data than the fine-tuning stage with a large variety of topics and speakers. The fine-tuning stage has the purpose of specifically adapting on selected datasets, in this case with the goal of conditioning the model on sentence embeddings.
LJSpeech and LibriTTS-R are used during the pre-training stage in addition to the emotional speech datasets ESD, RAVDESS and TESS in order to increase the overall quality of the models and to make them more robust against mispronunciations. However LJSpeech and LibriTTS-R are not suitable for testing the hypotheses of this thesis since

the prosodic parameters are not labeled in any way, i.e. no emotion labels are provided. Furthermore it cannot be assumed that there is a strong correlation between the meaning of the utterances and their prosodic realization.
During the fine-tuning stage only the emotional speech datasets are used such that the model is exposed to a high correlation between prompt embeddings and speech prosody.

For the proposed model, prompt embeddings are extracted from the emotional text dataset Yelp such that the emotion labels from text and speech correspond to each other. The sentence embeddings are selected randomly at each training iteration such that every speech sample is linked with a lot of different prompts. This strategy has the advantage that a large number of different prompts is seen during training which reduces the risk of over-fitting and increases the generalization capabilities of the system. It is further beneficial for inference where it becomes possible to provide arbitrary natural language inputs in order to control the prosody of generated speech. During the pre-training stage sentence embeddings for LJSpeech and LibriTTS-R have to be provided as well. Since these datasets are not labeled with emotion categories, sentence embeddings of the utterances themselves are extracted and used for conditioning. While there is no strong correlation between the utterances and their prosodic realization, the sentence embeddings might still be beneficial for improving overall speech quality since due to the contained semantic and syntactic information. Aside from that LJSpeech and LibriTTS-R are not used during the fine-tuning stage, giving the model the possibility to adapt to the strong correlation present between the emotional speech datasets and selected prompt embeddings.

The TTS models consist of roughly 46,000,000 trainable parameters and have a hidden dimensional size 192. The implementation and corresponding training code can be accessed on GitHub [26]. A demo page containing several audio samples is available as well [27]
The hyper-parameters of both stages are identical except of the amount of training steps which is 120,000 for pre-training and 80,000 for fine-tuning. Adam (Kingma and Ba, 2014) is used as optimizer with a learning rate of 0.001. The amount of warm-up steps, i.e. with custom learning rate, is 8000 and the joint training of the flow-based post-net is started after 9000 steps. The models were trained with a batch size of 32 on a NVIDIA-RTX-A6000 GPU.
Training charts for each individual loss during fine-tuning can be seen in figures 8-11. In

[26] https://github.com/Thommy96/IMS-Toucan

[27] https://thommy96.github.io/toucanpromptingdemo/

each case the loss of the proposed system is smaller than for the baseline which gives preliminary insights with regard to the usefulness of prompt embeddings. After training was complete, the overall loss for the baseline was 163.40 compared to 107.32 for the proposed system.

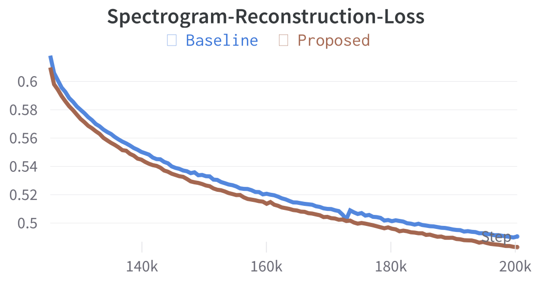


Figure 8: Spectrogram-Reconstruction-Loss

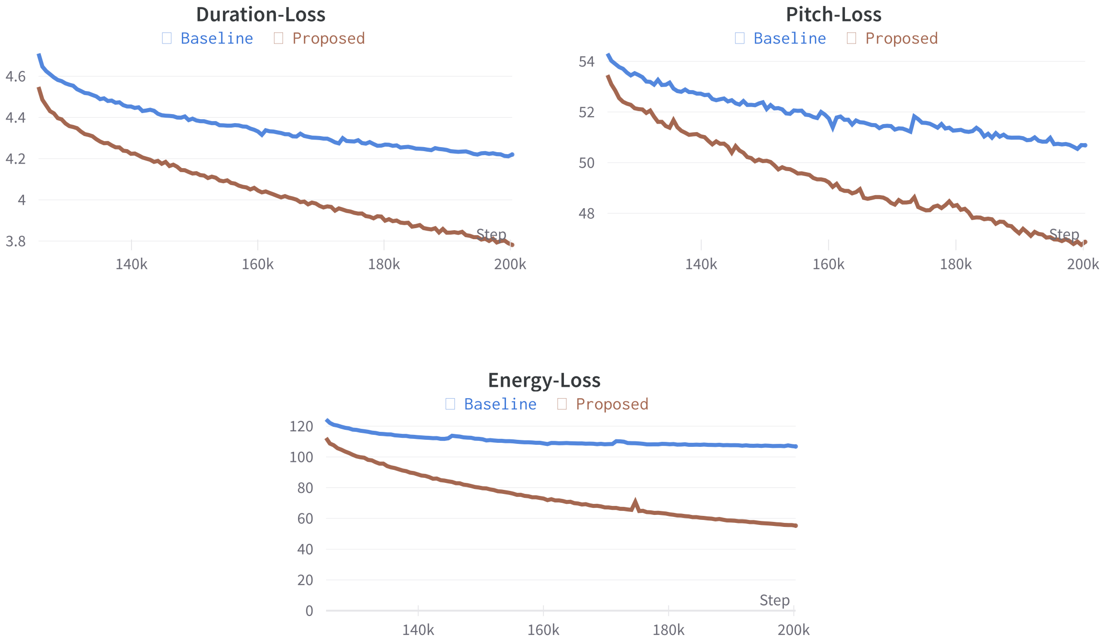


Figure 9: Prosody-Predictor-Losses

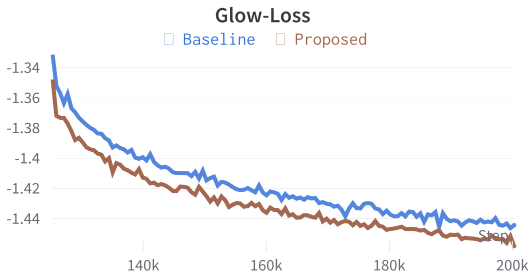


Figure 10: Glow-Loss

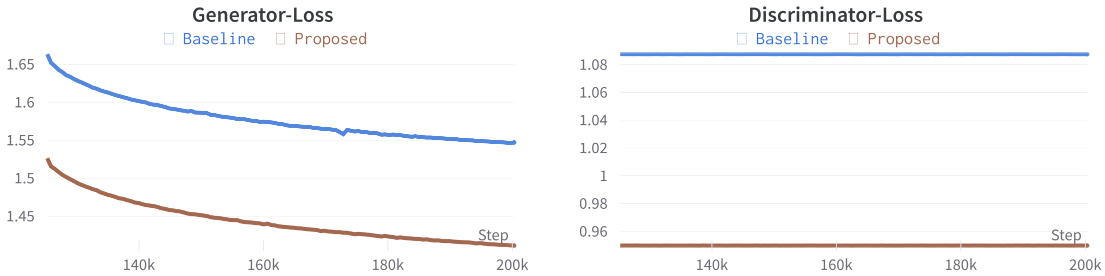


Figure 11: Adversarial-Losses

5.5 Inference Procedure

For evaluation purposes, sentences from the emotional text datasets DailyDialog (Li et al., 2017) and Tales-Emotion (Alm et al., 2005; Alm and Sproat, 2005) with annotated emotions *anger*, *joy*, *neutral*, *sadness* and *surprise* are synthesized. The emotional states *disgust* and *fear* are not considered since they don't appear in the ESD dataset on which the evaluation is based. Furthermore, sentences where the annotated emotion label doesn't match with the prediction of the DistilRoBERTa emotion classifier are discarded. Then the top 50 sentences per emotion category for each dataset according to the probability score of the emotion classifier are used for synthesis. This ensures that only sentences with high emotional content from each category are taken into account and reduces the risk of including inconclusive or ambiguous instances. The total number of test sentences for objective evaluation is 300 per dataset. For subjective evaluation, one sentence from each emotion category of each dataset is selected manually, resulting in 10 test sentences.

All test instances are synthesized using the baseline model as well as the proposed model. When using the proposed model the prompt which is provided to the sentence embedding extractor and is further used to condition the prosody of generated speech can be chosen independently from the input text.
In subjective experiments each test sentences is synthesized twice with the proposed model, the first time using the sentence itself as prompt and the second time using a sentence of a different emotion category. The speaker identities of a female speaker (id 15 in ESD) and a male speaker (id 14 in ESD) are selected for synthesis. The samples for subjective evaluation along with their mapping to other sentences can be seen in table 3.
In objective evaluation each of the 50 sentence belonging to an emotion category is additionally synthesized using a sentence from every other emotion category as prompt. By doing so it can be assessed if the expressed emotional states are transferred across sentences which might indicate if the prediction of the prosodic parameters relies on the provided sentence embedding.

The vocoder used in the experiments of this thesis is Avocodo (Bak et al., 2022) which has been found to produce the best sounding audio signals compared to the other vocoders during preliminary listening tests. The waveform signal is either saved in *wav* or *flac* file format to ensure lossless audio compression.

For all comparisons between ground truth and synthesized speech samples the effect of the vocoder has to be excluded since it can introduce additional noise or artifacts not present in the generated spectrogram. This is done by extracting the mel-spectrograms from ground truth speech samples and converting them back into waveform using the same vocoder as for synthesis. Furthermore this has the benefit that all audio signals have the same sampling rate and loudness.

Sentence	**Prompt**
You can't be serious, how dare you not tell me you were going to marry her?	You can go to the Employment Development Office and pick it up.
The king grew angry, and cried: That is not allowed, he must appear before me and tell his name!	Really? I can't believe it! It's like a dream come true, I never expected that I would win The Nobel Prize!
I really enjoy the beach in the summer.	He was astonished when he saw them come alone, and asked what had happened to them.
Then she saw that her deliverance was near, and her heart leapt with joy.	The sisters mourned as young hearts can mourn, and were especially grieved at the sight of their parents' sorrow.
You can go to the Employment Development Office and pick it up.	I really enjoy the beach in the summer.
So the queen gave him the letter, and said that he might see for himself what was written in it.	The king grew angry, and cried: That is not allowed, he must appear before me and tell his name!
Lily broke up with me last week, in fact, she dumped me.	You can't be serious, how dare you not tell me you were going to marry her?
The sisters mourned as young hearts can mourn, and were especially grieved at the sight of their parents' sorrow.	Then she saw that her deliverance was near, and her heart leapt with joy.
Really? I can't believe it! It's like a dream come true, I never expected that I would win The Nobel Prize!	So the queen gave him the letter, and said that he might see for himself what was written in it.
He was astonished when he saw them come alone, and asked what had happened to them.	Lily broke up with me last week, in fact, she dumped me.

Table 3: Test Sentences and Prompts used for Subjective Evaluation

Emotion Categories are indicated by text colors as follows:
anger, joy, neutral, sadness, surprise

5.6 Objective Evaluation Methods

5.6.1 Speaker Similarity

In order to verify that the TTS system has multi-speaker capabilities, the speaker similarity between speech samples from the training dataset and synthesized utterances is calculated. Thereby synthesized samples are generated with embeddings from the learned speaker embedding matrix corresponding to the speaker identity of the ground truth samples. The speaker similarity is computed as the cosine similarity between speaker embeddings from ground truth and synthesized speech. Thereby speaker embeddings are extracted using a pre-trained speaker verification model (Snyder et al., 2018) [28] provided by Speechbrain (Ravanelli et al., 2021).

The cosine similarity between two vectors $\overrightarrow{x}$ and $\overrightarrow{y}$ is determined by calculating the cosine of the angle between them as in equation 7.

$$\textit{cosine similarity} = \frac{\overrightarrow{x} \cdot \overrightarrow{y}}{\|\overrightarrow{x}\| \|\overrightarrow{y}\|} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2 \cdot \sum_{i=1}^{n} y_i^2}} \tag{7}$$

The cosine similarity ranges between -1 and 1 where 1 indicates perfect similarity and -1 high dissimilarity.

5.6.2 Word Error Rate

As a measure of intelligibility the word error rate of an automatic speech recognition (ASR) system is calculated. Therefore, the synthesized waveforms are first transcribed with Facebook's pre-trained Wav2Vec2 model (Baevski et al., 2020) [29]. Then the word error rate (WER) between ground truth sentences from the test datasets and transcribed utterances is calculated as in equation 8. It indicates the percentage of words which are incorrectly predicted under consideration of the number of substitutions S, deletions D and insertions I that have to be applied to the transcribed sentence in order to restore the ground truth. The sum of those is divided by the total number of words N.

$$\textit{WER} = \frac{S + D + I}{N} \tag{8}$$

[28] `https://huggingface.co/speechbrain/spkrec-xvect-voxceleb`

[29] `https://huggingface.co/facebook/wav2vec2-base-960h`

While a small word error rate indicates high intelligibility it has to be noted that the computed value can theoretically be greater than 100% since the sum of substitutions, deletions and insertions might exceed the number of words in the ground truth sentence.

5.6.3 Emotion Recognition

In order to objectively assess the capability of the TTS system to generate speech expressing an emotional state corresponding to the provided sentence embedding, the expected emotion category is compared to the prediction of an auxiliary speech emotion recognition model. In the experiments of this thesis an emotion recognition model by SpeechBrain (Ravanelli et al., 2021) [30] is applied. However their model is pre-trained on the IEMOCAP dataset (Busso et al., 2008) and performs poorly when applied to other datasets such as ESD. This might be caused by the fact that emotions are expressed differently across speakers. Furthermore the recording environment and the approach to trigger emotional states in speakers might differ. Aside from that the emotion labels available in IEMOCAP do not match the ones in ESD. Therefore, for the purposes of this thesis the speech emotion recognition is trained from scratch on the ESD dataset with the training code provided by SpeechBrain. The corresponding code can be accessed on GitHub [31]. The model is then used to predict emotion categories of synthesized test sentences which are further compared to the underlying emotion categories of provided sentence embeddings by calculating the accuracy as in equation 9.

$$\text{(9)} \qquad Accuracy = \frac{number\ of\ correctly\ predicted\ labels}{number\ of\ sentences}$$

[30] https://huggingface.co/speechbrain/emotion-recognition-wav2vec2-IEMOCAP

[31] https://github.com/Thommy96/speechbrain

5.7 Subjective Evaluation Methods

The subjective evaluation is carried out through an online survey on Soscisurvey [32] where participants are asked to listen to speech samples and rate them mainly with respect to their prosodic properties and expressed emotion. Since it is not required to have any prior knowledge about the prosody of spoken language, an introductory explanation is given in the beginning of the questionnaire. Additionally it is explicitly pointed out that the prosody of speech samples is not necessarily appropriate to the content of the spoken sentence itself. For example the utterance could be *"I am happy."* although the prosody indicates that the speaker is sad. While this introduction should make the participants aware of rating the prosody independently from the text, it is still expected that the content of the utterance influences their judgment to some extent. The questions that were posed can be found in the appendix in section 10.1.

5.7.1 Mean Opinion Score

The mean opinion score (MOS) is a widely used measure applied to assess the perceived speech quality in terms of naturalness, fluency and intelligibility. Participants are asked to rate these properties on a 5-point scale ranging from *bad* to *excellent*. Mean opinion scores are collected for audio samples of each TTS system configuration, i.e. baseline, proposed model with prompt identical to the utterance and proposed model with prompt different to the utterance. Furthermore reconstructed ground truth speech samples were rated.

5.7.2 Conveyed Emotion

The participants are asked to listen to an audio sample which could stem from any TTS system configuration or ground truth speech. Then their task is to select one or multiple emotion categories which they think are conveyed by the speech prosody. Available emotion categories were the ones present in the ESD dataset, i.e. *anger*, *neutral*, *surprise*, *joy* and *sadness*.

[32] `https://www.soscisurvey.de/`

5.7.3 Valence-Arousal Rating

According to the continuous circumplex model (Russell, 1980), emotions can be described by the two dimensions valence and arousal. Participants are asked to rate these two properties on 5-point scales respectively whereby valence ranges from *not pleased (negative)* to *pleased (positive)* and arousal from *calm (low arousal)* to *excited (high arousal)*. This rating task is included additionally to the selection of conveyed emotions since it provides a more fine-grained assessment of speech prosody. Furthermore it might be easier for the listeners to give ratings than to decide between discrete categories.

5.7.4 A/B Preference Test

Participants are asked to listen to two audio files, one generated by the baseline and the other one by the proposed system while conditioning the model on the sentence embedding of the utterance itself. The audio files are presented in randomized order to reduce positional effects. Then the task is to select the preferred audio file with respect to its prosodic appropriateness to the emotional content of the utterance. It is also possible to indicate that none of the both audio files is preferred.

5.7.5 Prosodic Similarity Test

This test has aims to assess if the a specific prompt results in similar prosody when used for the generation of different utterances. Therefore, participants are asked to listen to two audio files which are both synthesized by the proposed system using the same sentence embedding as conditioning signal. They were then obliged to rate how similar the samples sound with respect to their prosodic realizations. The ratings are collected on a 5-point scale ranging from *not similar at all* to *very similar*.

6 Results

6.1 Objective Evaluation

6.1.1 Speaker Similarity

The box-plots in figure 37 in the appendix show the distribution of calculated speaker similarities with respect to ground truth speech samples. For each speaker, outliers are discarded for further evaluation. As can be seen in table 4, the mean speaker similarity scores are consistently high (>0.9) for all speakers regarding both the baseline and the proposed system. The overall average score is nearly identical for both systems (baseline: 0.950, proposed: 0.953).

Speaker	**Baseline**	**Proposed**
11	0.951	0.954
12	0.948	0.953
13	0.954	0.956
14	0.951	0.955
15	0.951	0.954
16	0.949	0.951
17	0.938	0.948
18	0.952	0.955
19	0.948	0.950
20	0.951	0.954
Overall	0.950	0.953

Table 4: Speaker Similarity

6.1.2 Word Error Rate

The data distribution of calculated word error rates for all speakers is presented in the box-plots in figure 38 (appendix). Table 5 provides an overview of the mean word error rates across systems and speakers. It can be observed that ground truth speech almost always exhibits the lowest word error rates, especially for speakers 12, 13, 15, 16 and 18 where it even reaches 0. However, for speakers 11 and 17 the word error rate is slightly higher than the baseline and proposed system (0.121 and 0.093 respectively). Furthermore, the baseline generally outperforms the proposed system except for speaker 13 (0.076 compared to 0.072). Universal scores, i.e. averaged over all speakers, are visualized in figure 12 and reveal the lowest word error rate for ground truth speech (0.038) followed by utterances generated by the baseline (0.069) and the proposed system (0.078).

Speaker	**Ground Truth**	**Baseline**	**Proposed**
11	0.121	0.111	0.105
12	0.000	0.055	0.071
13	0.000	0.076	0.072
14	0.054	0.067	0.081
15	0.000	0.058	0.065
16	0.000	0.062	0.088
17	0.093	0.065	0.076
18	0.000	0.054	0.073
19	0.072	0.072	0.079
20	0.041	0.069	0.087
Overall	0.038	0.069	0.078

Table 5: Word Error Rates

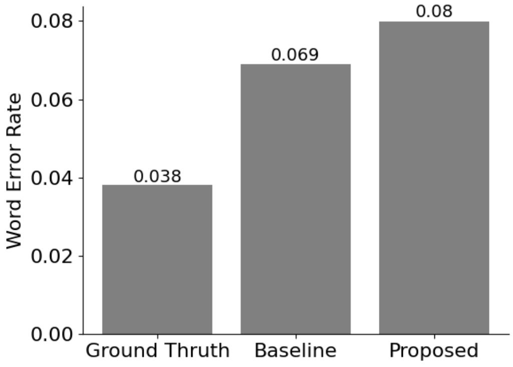


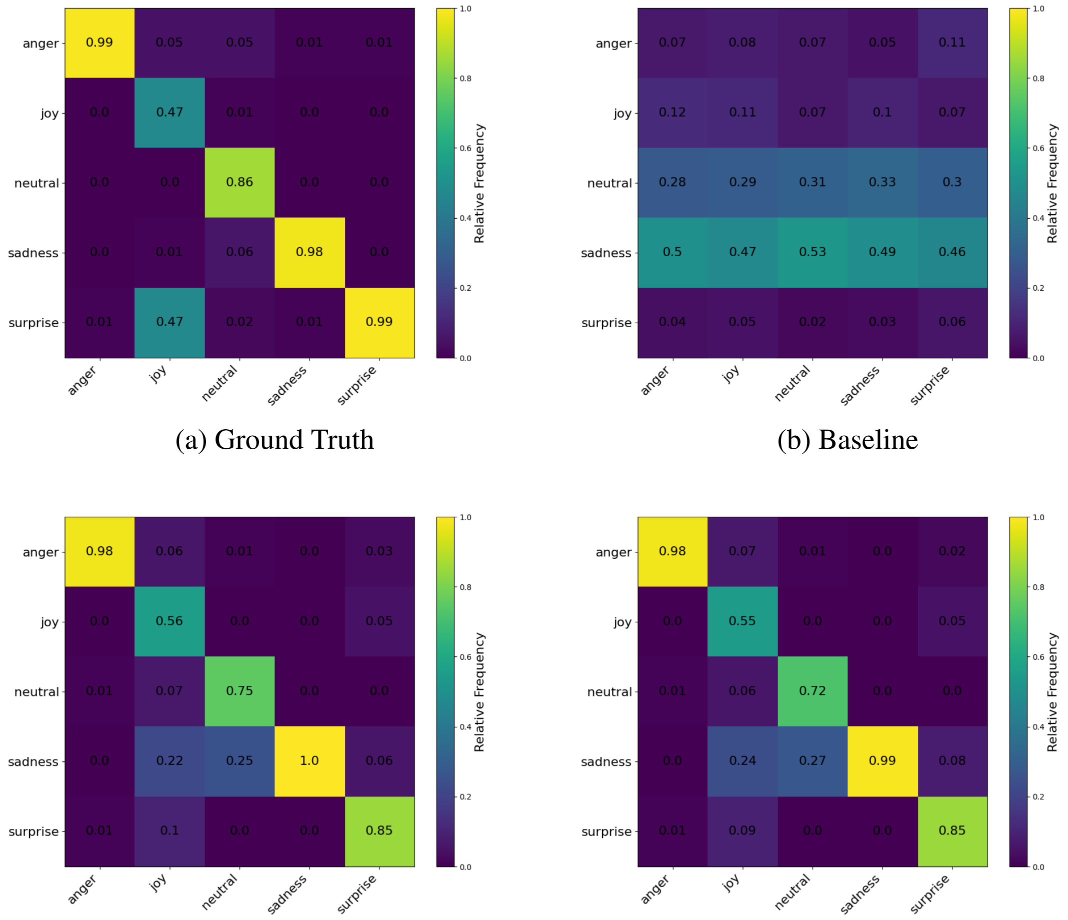
Figure 12: Overall Word Error Rates

6.1.3 Emotion Recognition

The emotion recognition model's predictions for all test sentences across different speakers are grouped based on emotion categories. The confusion matrices in figure 13 illustrate the relative frequencies of predicted emotion labels opposed to actual underlying emotion categories.
For ground truth speech these underlying emotions are given by the annotations in the ESD dataset. Figure 13a indicates that the model achieves high accuracy in predicting emotions for speech samples corresponding to anger (0.99), neutral (0.86), sadness (0.98) and surprise (0.99). However, joyful speech is often misclassified as surprise (0.47).
Considering the baseline, the underlying emotions correspond to the emotional content of the synthesized sentences as given by the annotation in the emotional text datasets. As can be seen in figure 13b speech samples of all underlying emotions are most often predicted to belong to the sadness category (anger: 0.5, joy: 0.47, neutral: 0.53, sadness: 0.49, surprise: 0.46). Apart from that, the category neutral is predicted second most frequently for all underlying emotions (anger: 0.28, joy: 0.29, neutral: 0.31, sadness: 0.33, surprise: 0.3).
With regard to the proposed system, the underlying emotions represent the emotional content of the provided conditioning prompt based on the annotation in the emotional text datasets. The results are further divided into cases where the prompt is identical to the input text (Proposed Same) and cases where another prompt is used (Proposed Other). Figures 13c and 13d show the relative frequencies of predictions for these two cases. Generally the underlying emotion of the prompt is also predicted in synthesized speech with an especially high accuracy for anger (Proposed Same / Proposed Other: 0.98) and sadness for which almost prefect scores can be observed (Proposed Same: 1.0, Proposed Other: 0.99). Speech samples generated with a neutral prompt are also mostly predicted as neutral with relative frequencies of 0.75 (Prompt Same) and 0.72 (Prompt Other). The scores for joy are substantially lower compared to the other emotion categories, however the prediction still aligns for more than half of the test sentences (Proposed Same: 0.56, Proposed Other: 0.55).

In order to assess the association strength between underlying and predicted emotions, Cramér's V (Cramér, 1999) is calculated. This measure is based on Pearson's chi^2 statistic which compares observed frequencies to expected ones under the assumption of independence between the variables. It has to be noted that common correlation metrics such as Pearson's r or Spearman's ρ are not applicable since the variables are nominal and not sortable on a scale. Cramér's V ranges from 0 and 1 whereby higher values indicate a higher association strength. The respective values can be found in table 6 revealing high association strengths for ground truth speech (0.85) as well as for the proposed system (0.80). The baseline immensely lacks behind with a value of 0.06. All calculated scores in the table have found to be significant given a significance level of 0.05.

Finally, figure 14 illustrates the overall accuracy, i.e. the proportion to which the underlying and predicted emotion categories align. Thereby high scores can be observed for ground truth speech (0.86) as well as for the proposed system (Proposed Same: 0.83, Proposed Other: 0.82) whereas the baseline only achieves an accuracy of 0.21.



(a) Ground Truth

(b) Baseline

(c) Proposed Same

(d) Proposed Other

Figure 13: Predicted Emotions

	Ground Truth	Baseline	Proposed Same	Proposed Other
Cramér's V	**0.85**	**0.06**	**0.80**	**0.80**

Table 6: Cramer's V between underlying and predicted emotion categories

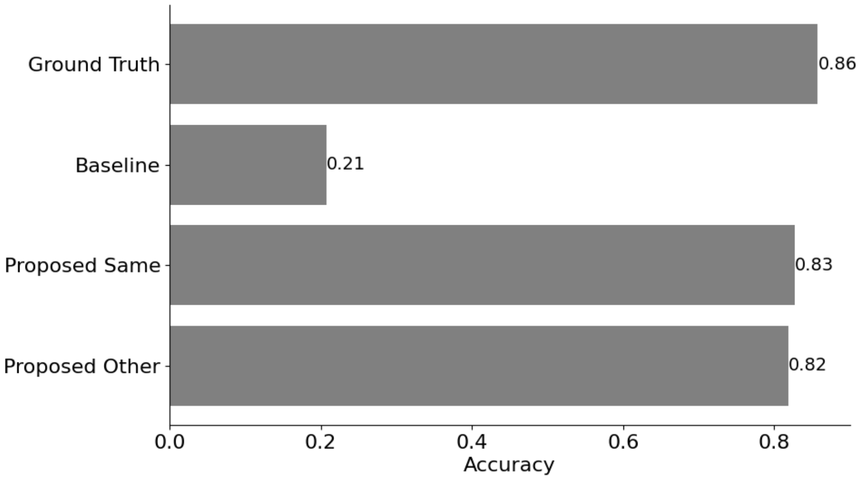
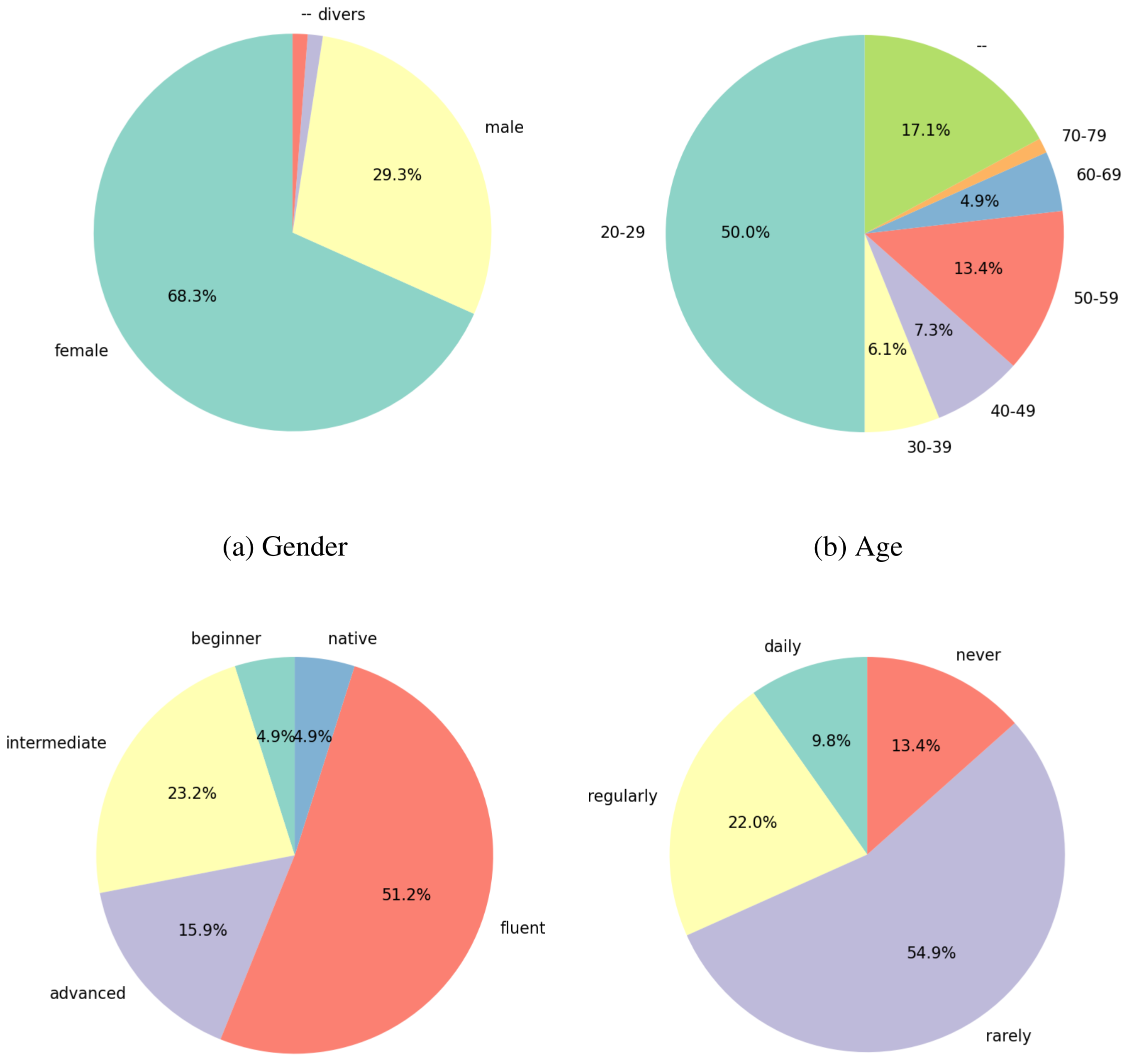


Figure 14: Overall Accuracy

6.2 Subjective Evaluation

6.2.1 Sociodemographics

Sociodemographic information such as gender, age and the level of English skill was collected previous to the listening tests. Additionally the experience with synthetic speech has been assessed by asking how often a participant listens to artificially generated speech, e.g. by using voice assistants. All sociodemographic questions were non-obligatory.
In total 82 people participated in the study, figure 15 gives insights about the distribution of this sample. There are more than twice as much female (68.3%) than male participants (29.3%). Furthermore, half of them are between 20 and 29 years old and more than half (51.2%) are fluent English speakers. 54.9% of participants state that they rarely listen to synthetic speech whereas 31.8% have considerable experience by listening to it regularly or daily.



(a) Gender

(b) Age

(c) English Skills

(d) Listening to Synthetic Speech

Figure 15: Sociodemographics

6.2.2 Mean Opinion Score

Box-plots with Mean Opinion Scores for each system across speakers and emotions can be seen in figures 33-35. They indicate how often a specific rating is given and visualize the distribution of the data. Outliers which are outside 1.5 × the interquartile range are discarded from further evaluation.

Table 7 shows accumulated Mean Opinion Scores for each system, speaker and emotion. In most cases the scores for the female speaker are higher than the ones for the male speaker, but there is no clear trend observable across emotions. Furthermore, for both speakers, the overall highest scores can be found for ground truth speech (female: 4.10, male: 3.79), followed by the proposed system (female: 3.57, male: 3.18) and the baseline (female: 3.55, male: 3.05). This tendency is confirmed by figure 16 showing mean opinion scores averaged across speakers with 3.95 for ground truth, 3.30 for the proposed system and 3.37 for the baseline. However, it can be noted that for sad utterances of the female speaker the mean opinion scores for the proposed system (3.70) and the baseline (3.63) are slightly higher than the ones for ground-truth speech (3.47). While the proposed system overall receives a slightly better score than the baseline, an independent samples t-test indicates that this improvement is not significant (significance level: 0.05, *p-value*: 0.46). It has to be noted that an independent samples t-test is chosen instead of a paired samples t-test since a participant does not see every speech sample, i.e. the ratings are given by groups different people. Although these groups might overlap, there is no way to consistently extract paired ratings.

		Anger	**Disgust**	**Neutral**	**Sadness**	**Surprise**	**Overall**
Ground Truth	female	4.53	4.47	4.53	3.47	3.50	4.10
	male	3.63	4.25	4.0	3.44	3.65	3.79
Baseline	female	3.00	3.81	4.06	3.63	3.27	3.55
	male	3.45	2.87	3.44	3.06	2.44	3.05
Proposed	female	3.34	3.53	4.00	3.70	3.28	3.57
	male	3.09	3.53	3.33	3.06	2.88	3.18

Table 7: Mean Opinion Scores per Emotion

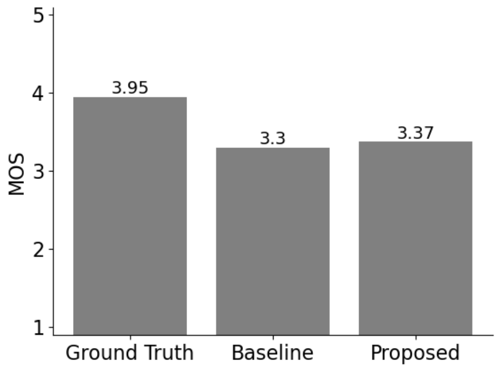


Figure 16: Overall Mean Opinion Scores

6.2.3 Conveyed Emotion

The confusion matrices in figures 17-20 visualize the emotions labeled by participants with respect to the real or intended underlying emotion. Thereby, each column in a matrix corresponds to the distribution of given emotion labels and each cell indicates the normalized frequency of a given emotion label within the respective column.

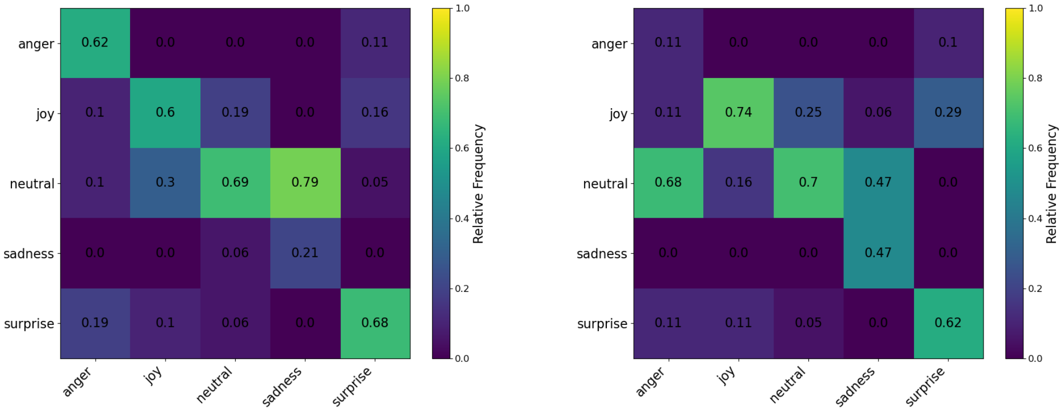
For ground truth speech (figure 17) the underlying emotion categories correspond to the ones as annotated in the ESD dataset. Notably, for both the female and the male speaker, high relative frequencies can be observed for joyful (female: 0.60, male: 0.74), neutral (female: 0.69, male: 0.70) and surprised (female: 0.68, male: 0.62) utterances being perceived as such. In most cases angry speech is rated accordingly for the female speaker (0.62) while it is predominantly perceived as neutral for the male speaker (0.70). Regarding the female speaker, sadness is very often perceived as neutral with a relative frequency of 0.79 compared to 0.21. Conversely, sad utterances from the male speaker have a relative frequency of 0.47 for both sad and neutral labels.

Moving to the baseline system (figure 18) the underlying emotion categories indicate the emotional content of the synthesized sentences themselves as annotated in the emotional text datasets. Generally, the utterances are mainly categorized as neutral. This tendency holds stronger for the female speaker than for the male speaker where joyful sentences are often noticed as being sad (0.56).

Nevertheless, in general, scores for the neutral category are still substantially high except for angry sentences which are often identified as such (0.4) or surprised (0.3). With regard to the female speaker, a substantial portion of sad sentences is judged as angry (0.39).

The results for the proposed TTS system are split based on whether the conditioning prompt matches the synthesized sentence (figure 19) or differs (figure 20). Hereby, the underlying emotion categories reflect the emotional content of the provided prompt as annotated in the emotional text datasets.
For the case where the prompt matches the input sentence, neutral utterances are perceived as such by all participants regarding the female speaker and by most participants (0.67) regarding the male speaker. Furthermore, for the male speaker the frequencies for all other emotion categories being neutral is also relatively high (anger: 0.45, joy: 0.55, sadness: 0.60, surprise: 0.35). However the scores on the diagonal are the second highest in those cases (anger: 0.32, joy: 0.25, sadness: 0.35), except for surprised utterances (0.30) which are more often judged as being joyful (0.35). Concerning the female speaker, one can observe a high frequency for joyful utterances being rated as neutral (0.82). Sentences belonging to the remaining emotion categories are mainly accordingly (anger: 0.70, sadness: 0.61, surprise: 0.44).
With regard to the scenario where the provided conditioning prompt differs from the synthesized sentence, a high relative frequency is evident for utterances where the prompt is joyful but judged as sounding neutral (female: 0.94, male: 0.38). Utterances from the female speaker with angry and neutral prompts are mainly perceived as such (0.80 and 0.48 respectively). For neutral and sad prompts this tendency also holds for the male speaker (0.67 and 0.68 respectively). However utterances with angry and joyful prompts are often rated as neutral (0.74 and 0.38 respectively) and surprised utterances have a relative frequency of 0.38 for anger and 0.32 for neutral compared to 0.33 for surprise. Regarding the female speaker, utterances with sad prompts are often perceived as neutral (0.50 compared to 0.30) and utterances with surprised prompts are rarely judged as surprised (0.19) but instead as joyful (0.38) or neutral (0.31).

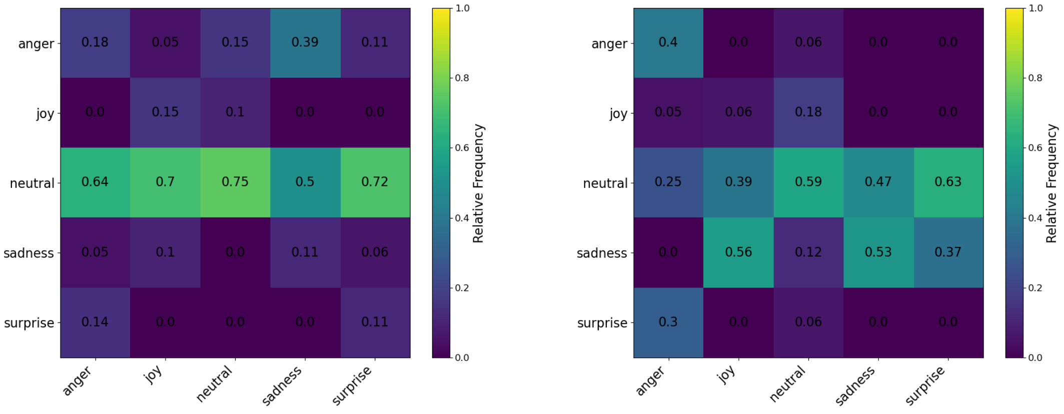
The overall association strength between underlying emotion categories, e.g. indicated by the emotional content of prompts, and conveyed emotions in speech prosody, as labeled by the participants, is measured by calculating Creamér's V (Cramér, 1999). Scores can be seen in table 8 with significant values in bold given a significance level of 0.05. All association strengths are significant except for the baseline's female speaker (0.27). The scores for ground truth speech are 0.56 for the female speaker and 0.53 for the male speaker. The proposed system for the female speaker, using the same prompts as the synthesized sentences, shows the overall highest value (0.61). All remaining values are below 0.5 but nevertheless significant, indicating a fair amount of association strength between the variables.



(a) female

(b) male

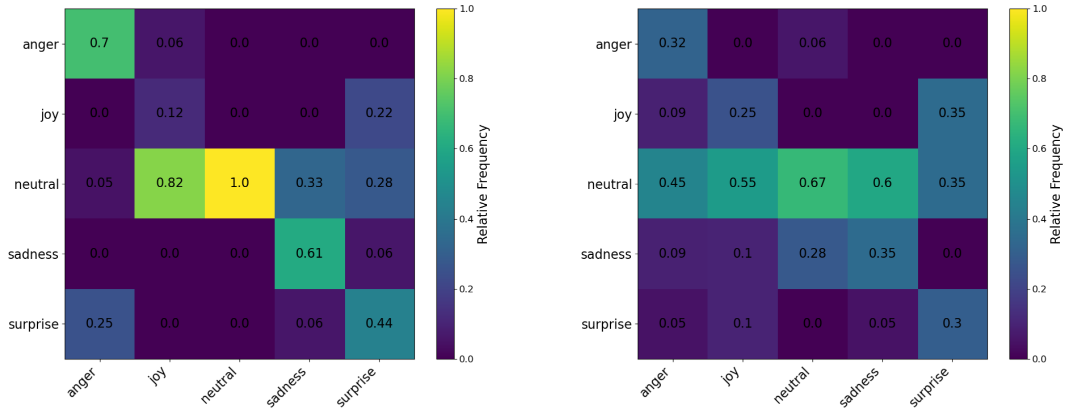
Figure 17: Conveyed Emotions, Ground Truth



(a) female

(b) male

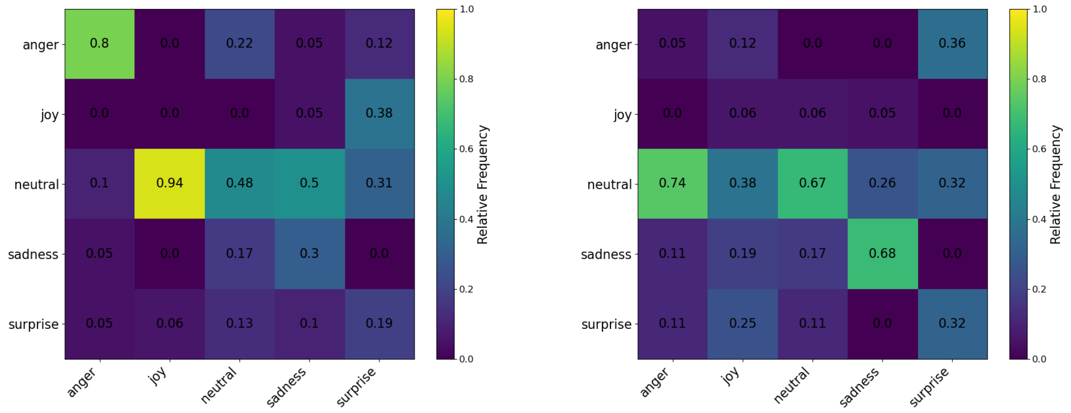
Figure 18: Conveyed Emotions, Baseline



(a) female

(b) male

Figure 19: Conveyed Emotions, Proposed Same



(a) female

(b) male

Figure 20: Conveyed Emotions, Proposed Other

		Ground Truth	**Baseline**	**Proposed Same**	**Proposed Other**
Cramér's V	female	**0.56**	0.27	**0.61**	**0.47**
	male	**0.53**	**0.43**	**0.38**	**0.40**

Table 8: Cramer's V between underlying and conveyed emotion categories

6.2.4 Valence-Arousal Ratings

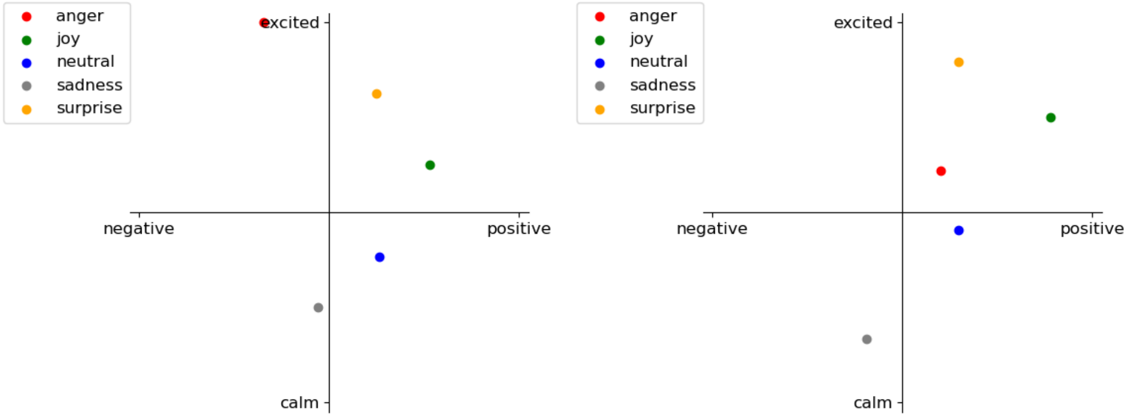
Accumulated valence and arousal ratings for each system, speaker and emotion, after removing outliers, can be seen in tables 9 and 10 in the appendix.
Figures 21-24 visualize these ratings using the continuous circumplex model. Each point in a plot corresponds to the mean valence and arousal value for the respective underlying emotion.

For ground truth speech 21 the underlying emotion relates to the emotion category as labeled in the ESD dataset. It can be observed that ratings for joyful, neutral, sad and surprised utterances are similar distributed for both speakers. For angry utterances there is a substantial difference in given ratings across speakers. For the female speaker such utterances are rated with very high arousal (5.00) and relatively low valence (2.31) while for the male speaker they are rated with much less arousal (3.44) and higher valence (3.40).

Regarding the baseline system 22, the underlying emotions correspond to the emotional content of the synthesized sentences as annotated in the emotional text datasets. The plots demonstrate that the points are generally grouped closer together and more centered around the origin. This tendency holds stronger for the female speaker as for the male speaker where specifically the ratings for surprise are more extreme (valence: 2.29, arousal: 1.71). It can also be noted that there the arousal ratings for both speakers don't exceed 3.5.

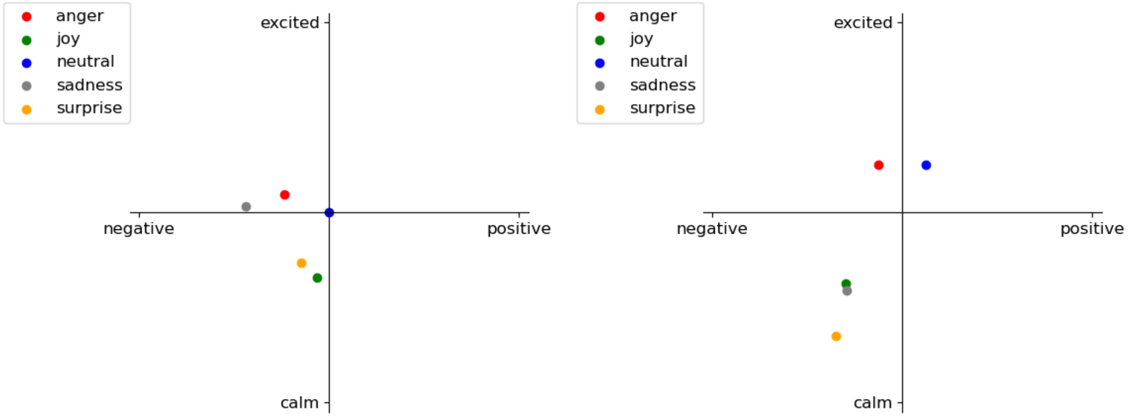
Concerning the proposed system where the conditioning prompts match the synthesized sentences 23, the underlying emotions correspond to the emotional content of the prompts as annotated in the emotional text datasets. In general, the distribution of ratings for the female speaker has considerable similarity with the one for ground-truth speech. Thereby, angry utterances have a high arousal (4.40) and a low valence ratings (1.44) while sad utterances are rated with low arousal (1.80) and low valence (1.88). For the male speaker, angry utterances receive a comparably low arousal score (2.40). Generally, the ratings for the male speaker are less extreme as for the female speaker.

With regard to the proposed system where the conditioning prompts are different from the synthesized sentences, the underlying emotions also correspond to the emotional content of the prompts. For the female speaker, utterances produced with angry prompts are rated with high arousal (4.53) and low valence (1.65) while obtaining more average scores for the male speaker (arousal: 3.38, valence: 3.00). Utterances generated with surprised prompts receive high arousal ratings for both speakers (female: 4.21, male: 4.00), while valence for the female speaker (3.67) is substantially higher than for the male speaker (2.47). Utterances with joyful prompts receive relatively moderate values (female: arousal: 3.19, valence: 3.00; male: arousal: 3.50, valence: 2.67) while neutral and sad utterances are judged with low arousal (female: 2.47 and 2.40, male: 1.94 and 1.94 respectively) and low valence (female: 2.06 and 2.59, male: 3.40 and 2.29 respectively).



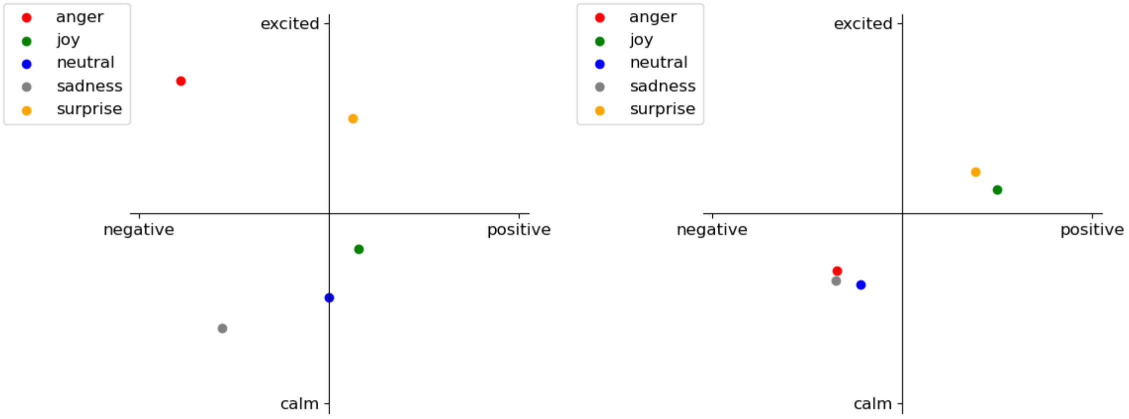
(a) female (b) male

Figure 21: Valence-Arousal, Ground Truth



(a) female (b) male

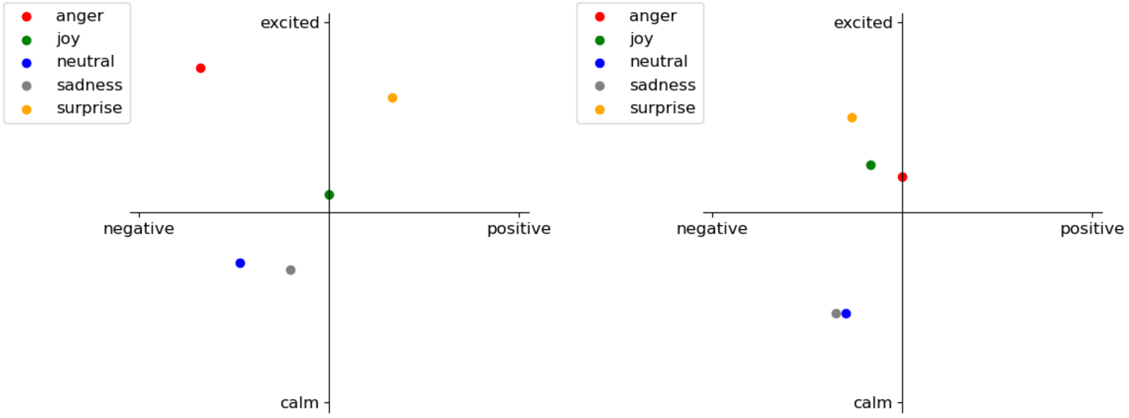
Figure 22: Valence-Arousal, Baseline



(a) female

(b) male

Figure 23: Valence-Arousal, Proposed Same



(a) female

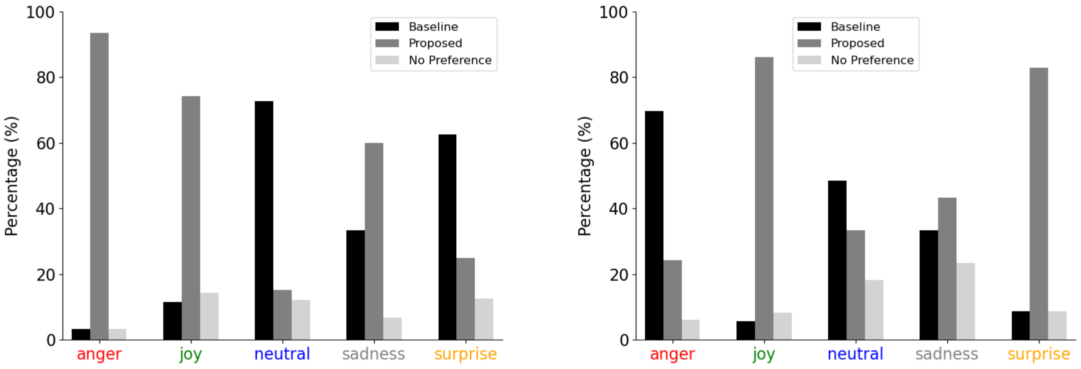
(b) male

Figure 24: Valence-Arousal, Proposed Other

6.2.5 A/B Preference

Figure 25 shows the distribution of preferences between the baseline and the proposed TTS system across emotion categories. The underlying labels correspond to the emotional content of the synthesized sentence which is also used as conditioning prompt in the proposed system.
With respect to the female speaker, the proposed system is clearly preferred for angry, joyful and sad utterances while the baseline is favored for neutral and surprised sentences. Regarding the male speaker, the proposed system is preferred for joyful and surprised utterances. For sad sentences the proposed system is slightly favored over the baseline which is clearly preferred for angry and moderately for neutral sentences.

An overall preference of the proposed system can be observed for both the male (53.42% compared to 36.65%) as well as the female speaker (55.09% compared to 32.34%). This tendency is visualized in figure 26 which shows that in total the proposed system is favored in 54.27% of cases compared to 34.45% regarding the baseline.



(a) female

(b) male

Figure 25: Preference across Speakers and Emotions

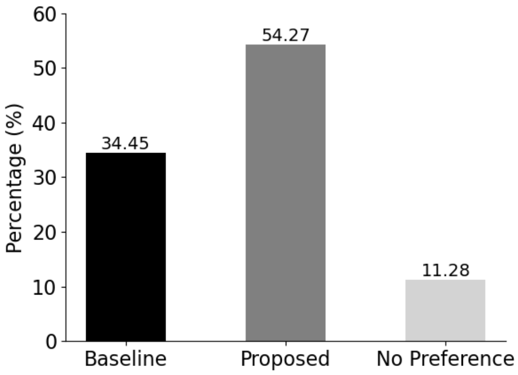
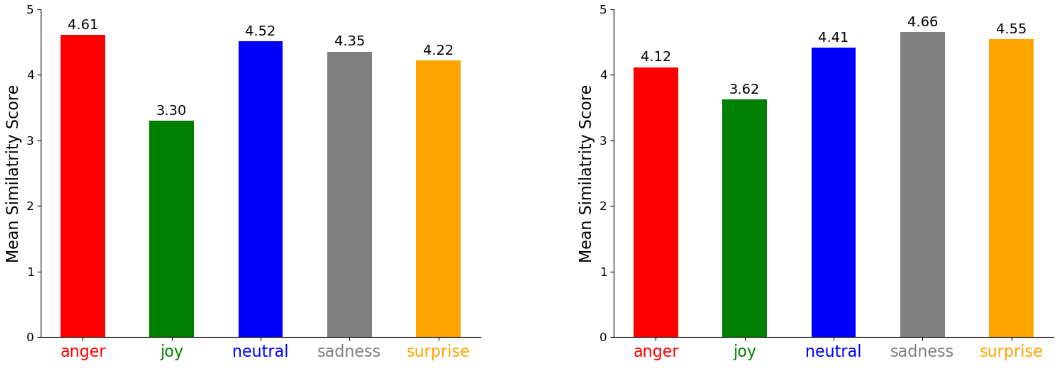


Figure 26: Overall Preference

6.2.6 Prosodic Similarity

The distribution of ratings given by the participants regarding the prosodic similarity between two different utterances generated by the proposed system while using the same conditioning prompt can be seen in the box-plots 36a and 36b in the appendix. Outliers are removed from the data for further evaluation.

Figure 27 shows averaged similarity scores across emotions for each speaker. Thereby, the underlying emotions correspond to the emotional content of the provided prompts. It can be observed that all similarity scores are very high (≥ 4.12) except for utterances generated with joyful prompts. However, in these cases the scores are still considerably high (female: 3.30, male: 3.62). The total mean similarity scores averaged over emotions as visualized in figure 28 confirm the high ratings (female: 4.20, male: 4.27).



(a) female

(b) male

Figure 27: Mean Similarity across Speakers and Emotions

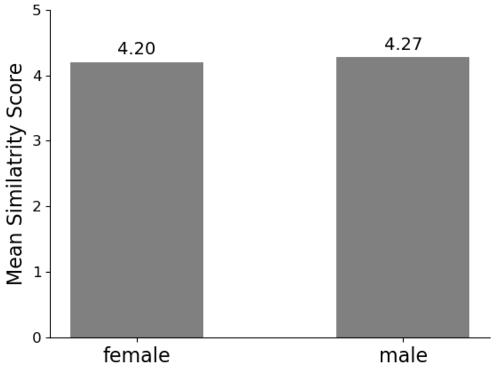


Figure 28: Overall Mean Similarity

7 Discussion

7.1 Quantitative Analysis

7.1.1 Multi-Speaker Capabilities

The high speaker similarity scores with respect to ground truth speech indicate that the speaker identities are almost perfectly preserved during synthesis. This proves the effectiveness of using a speaker embedding look-up table which learns the embeddings during training. Furthermore, it is in line with findings of previous work that follow the same approach for multi-speaker TTS (Ping et al., 2017; Chen et al., 2020). More interestingly, high speaker similarity is maintained in the proposed system demonstrating that the joint vector representation of speaker and emotional content retains the required information and is therefore a suitable conditioning signal for TTS. It can be hypothesized that this is specifically achieved by the squeeze and excitation bottleneck which helps modeling inter-dependencies between the two sources. Such a bottleneck has also been successfully applied in Pan and He (2021) for disentangling prosody and speaker timbre. While the integration of a squeeze and excitation block has been found to be useful in preliminary listening tests, it is suggested that this observation should be confirmed by ablation studies in future research.

Thus, these results suggest that hypothesis 5 can be accepted. The controllability of speaker identities is not affected by the integration of sentence embeddings.

7.1.2 Speech Quality

The low word error rate for ground-truth speech samples shows a high intelligibility, as expected. Although the score for the baseline system is comparably higher, it can still be interpreted as producing highly intelligible speech. The proposed system exhibits only a slightly higher error rate than the baseline, indicating that the quality is generally maintained. In the same vein, the mean opinion score from the subjective evaluation shows that speech quality and naturalness, only approximately 0.6 rating points lower than ground truth speech, is obtained by the proposed and baseline system. Hereby, it has to be noted that ground truth speech also includes potential quality degradation introduced by the vocoder while original recordings would probably achieve even higher mean opinion scores. However, converting original speech to mel-spectrogram and transforming it back into waveform with the vocoder allows for better comparison with synthesized samples. Under these considerations the TTS systems achieve state-of-the-art fidelity with only relatively small degradation compared to human speech. Moreover, these findings follow the tendency of related work based on similar model architectures (Ren et al., 2020; 2021; Lancucki, 2021). While the mean opinion score for the proposed system is marginally higher than for the baseline, this difference has not been found to be significant. However, it is important to mention that the evaluation is based only on a relatively small number of samples. Ratings for more test sentences would be required to obtain more conclusive results.

Nevertheless, the first hypothesis has to be rejected. Conditioning the TTS model on sentence embeddings of prompts does not improve the overall quality of generated speech.

7.1.3 Prosodic Controllability

The emotion recognition model shows very high accuracy for the prediction of emotion labels in ground truth speech on which it was trained, except for joy, meaning that the emotion can generally be recognized reliably from speech. Considering this, the strong alignment between underlying and predicted emotion labels for the proposed system demonstrates that the emotional content of the prompt is accurately transferred to speech, indicating a high degree of controllability. Moreover, the speech prosody exclusively relies on the provided prompt and is not influences by the input text of the synthesized utterance as revealed by the high accuracy when mixing prompts with input texts belonging to different emotion categories. Another remarkable observation is that in this scenario the accuracy for joy is even higher than for ground truth speech which implies that the proposed model mostly generates utterances where the conveyed emotion is clearly recognizable. In contrast, for the baseline the predicted emotion categories are mostly sadness and neutral, showing that there is hardly any prosodic variation in generated speech irrespective of the emotional content of the input text. These observations are further confirmed by the calculated Cramér's V values which demonstrate strong association between underlying emotions and predicted ones for the proposed system, but none for the baseline.

Results from emotion labeling and valence-arousal annotations by humans in the subjective evaluation for the female speaker allow similar conclusions, although not as evidently. In ground truth speech, the labeled emotions align well with the ones provided in the dataset. The only exception is sadness which is more often labeled as neutral. The distribution of valence-arousal ratings shows similar results with emotions being placed in space as could be expected, e.g. anger has a very high arousal value and negative valence. A comparable distribution can also be found in Russell (1980). The emotion labeling results regarding the baseline reveal that speech produced in the female speaker's voice is primarily perceived as neutral. This is further supported by the valence-arousal ratings which are much more centered around the origin. In contrast, for the proposed system a respectable alignment between provided prompts and perceived emotion in speech can be observed which is supported by the association strength marked by Cramér's V. Especially high agreement is present for anger as well as for neutral and sadness when using the same prompt as the input text. It is noticeably that speech generated with joyful prompts is most frequently judged as neutral which might indicate that there the degree of emotional intensity is quite low for this emotion. Another explanation could be that it is generally harder for participants to distinguish joyful from neutral speech compared to e.g. angry speech since the prosodic properties are more similar. Furthermore, when combining prompts and input texts from different emotion categories, the association strength decreases but is still found to be significant. A plausible explanation for this observation could be that although participants are asked to rate the speech emotion independently from the semantic content of the utterance, it probably still influences their decision since it is almost impossible to completely ignore it. Most remarkably in this scenario, speech produced using angry prompts is still most frequently judged accordingly signifying that the prosodic properties are modeled appropriately and are well distinguishable from the other emotion categories. The valence-arousal ratings for speech samples generated with angry prompts are rated even more extreme than for ground truth speech.

For the male speaker, human emotion labeling results and valence-arousal ratings are quite different to those of the female speaker. While the association strength for ground truth speech is the highest, the baseline emits a score similar to the proposed system. If looking at the confusion matrices for individual emotions it can be observed that synthesized speech for the male speaker is primarily judged as neutral or sad. Interestingly for ground truth speech, angry utterances are also mostly labeled as neutral and similar to the female speaker sadness is often judged as neutral as well. Considering the excellent results from emotion recognition supporting the tendency of results for the female speaker, it can be hypothesized that these observations might be a speaker specific issue. The male speaker might have a speaking style emitting prosodic properties that are generally perceived as neutral or sad. Additionally, it is possible that the emotions are not realized as intense as by the female speaker, causing the TTS model to produce more neutral sounding speech. Moreover, it has to be emphasized that in the subjective evaluations only one utterance from each speaker per emotion category is presented to the participants. Although these sentences have been selected manually, it might still occur that the results are influenced by sentence specific effects. More test sentences would be required to obtain more conclusive and reliable results.

Under consideration of the high accuracy scores from emotion recognition and the high association strengths of human emotion labeling for the female speaker in subjective evaluation, especially when compared to the baseline, it can be claimed that the proposed TTS system has profound prompting capabilities. The discussed results provide high evidence that the prosodic properties of generated speech can be accurately controlled by providing appropriate natural language prompts.

Furthermore, this conclusion is supported by the A/B preference test conducted in subjective evaluation. For more than half of the cases the proposed system is preferred over the baseline with respect to the adequateness of prosodic properties in generated speech. The baseline is only preferred in 34.45% of the cases whereby a large proportion of those are scenarios where neutral speech is expected, which is typical to be produced by the baseline.

In summary, the results indicate that hypotheses 2, 3 and 4 can be accepted. The prosodic parameters of speech and the conveyed emotional state are controllable by providing suitable prompts. Moreover, the emotional content of the prompts is highly correlated to the prosody and conveyed emotion of synthesized speech. Finally, semantically similar prompts, i.e. prompts of the same emotion category, result in similar prosody and conveyed emotion.

While there is recent work that also achieves prosodic controllability through natural language prompts, the work in this thesis follows an approach that is different in several aspects. Kim et al. (2021a); Shin et al. (2022); Guo et al. (2022b); Liu et al. (2023) and Yang et al. (2023) rely on datasets with manually provided annotations of style descriptions. It might be expensive and time consuming to collect such annotations, especially for large TTS datasets. Further, the obtained style descriptions might be restricted in their form and follow similar patterns which constrains the formulation of useful prompts by the user during inference. Approaches that solely rely on the input text to control the prosody (Wu et al., 2022; Mukherjee et al., 2022)) assume that the prosodic properties of the audio are related to the content of the utterance itself. While such a connection might be present in the audio-book domain to some extent, it cannot be assumed in general which has been observed by Mukherjee et al. (2022). Besides that, even in audio-books the prosody of an utterance might often depend on surrounding context sentences or the impersonation of specific characters by the voice actor.

The proposed approach avoids the construction of new datasets by combining existing emotional speech and text corpora, further obtaining high correlation between prompts and prosody which is crucial for TTS training. This technique is inspired by Tu et al. (2022) and extended in the aspect that appropriate prompts are selected randomly from a large pool which is expected to increase the generalization capabilities of the model since a lot more prompts are seen during training. Moreover by doing so, there is no risk of establishing connections between utterances and specific prompts. As a further distinction, the training strategy is applied to a TTS system with dedicated duration, pitch and energy predictors, potentially improving overall prosodic controllability. Finally, in contrast to Tu et al. (2022) which model speaker identities within the prompt, the proposed approach effectively combines prompt and speaker embeddings. Hence, this makes it a promising avenue for potential future research in the field of cross-speaker style transfer and zero/few-shot speaker modeling.

7.1.4 Emotional Style Transfer

The results from prosodic similarity tests, where a fixed prompt was used across different utterances, demonstrate very high scores. This suggests that the speech emotion caused by a prompt is effectively transferable to arbitrary utterances. Furthermore, these results demonstrate that the generated speech prosody heavily relies on the provided prompt while being completely unaffected by the input text itself. In addition, it could be seen as a possibility to clone the speaking style of one synthesized utterance and apply it to another. With regard to individual emotion categories, similarity scores for utterances produced with a joyful prompt are comparably lower. This might indicate that these prompts do not result in speech samples with well recognizable speaking styles or emotional states, making it harder for participants to judge them.

Nonetheless, these overall high similarity scores further suggest that hypothesis 4 can be accepted. Emotional Style Transfer is further investigated in the spectrogram case studies in section 7.2.2.

7.2 Qualitative Analysis

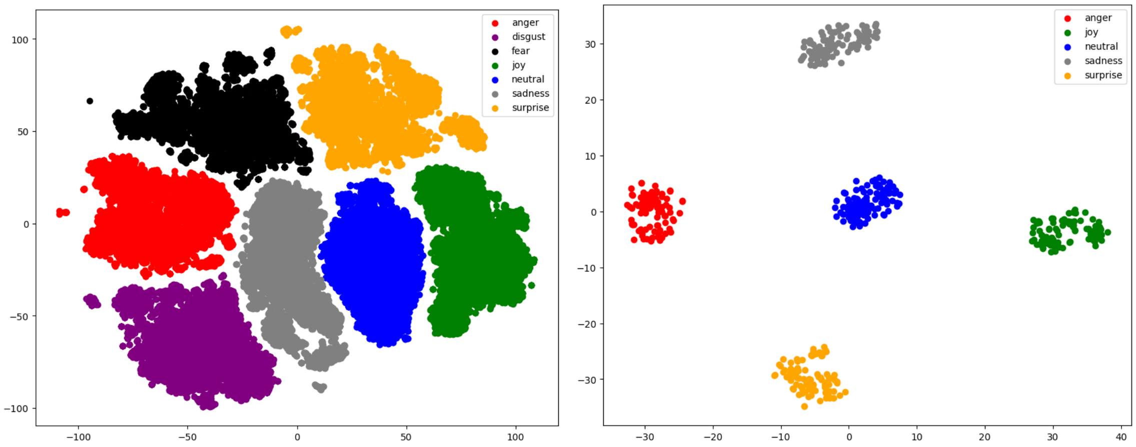
7.2.1 Adequateness of fine-tuned Sentence Embedding Extractor

The sentence embedding extractor as described in section 5.3 is claimed to be suitable for capturing the emotional content of sentences since it is fine-tune on the task of emotion prediction. The usefulness of extracted embeddings is further investigated by producing t-distributed stochastic neighbor embedding (t-sne, Hinton and Roweis (2002)) plots. Figure 29a shows sentence embeddings extracted from emotional prompts used during training as points in a 2-dimensional space. Figure 29b analogically visualizes the test sentences used for evaluation. The color of each point in the figures indicates the corresponding emotion label. It can be seen that the sentence embeddings build clusters in the embedding space which are well separable, verifying their usefulness. In contrast, sentence embeddings extracted from the *[CLS]* token of a standard BERT model (Devlin et al., 2019) (figure 30a) as well as embeddings produced by a sentence transformer model (Reimers and Gurevych, 2019) (figure 30b) have mixed up distributions with respect to the emotion labels. This demonstrates that they do not capture emotional content per se.

7.2.2 Spectrogram Case Studies

Spectrogram case studies are presented in order to further investigate the influence of prompts for the duration, pitch and energy predictors as well as the emotional style transfer achievable through prompting. The speaker identity of the female speaker which is also used for subjective evaluation is chosen for this purpose.

Figure 31 illustrates spectrograms with corresponding pitch contour in red for two content-wise neutral utterances generated with identical emotional prompts for anger and surprise respectively. It can be seen that the produced pitch curves are very similar for both sentences. Regarding anger, pitch goes moderately down at the end of the sentence whereas for surprise it increases drastically making the sentence sound like a surprised question.

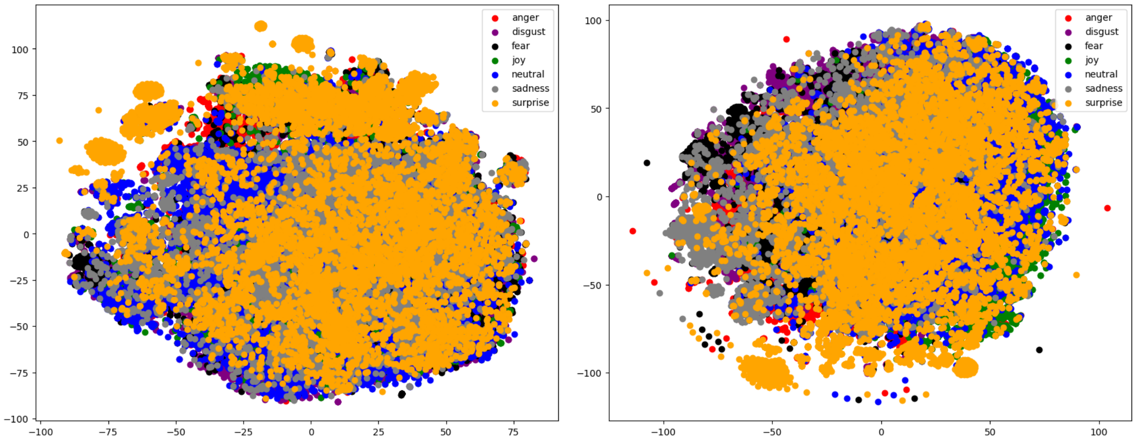


(a) Prompts used for Training

(b) Prompts used for Evaluation

Figure 29: T-SNE Plots of applied Sentence Embeddings

Extracted with the RoBERTa model fine-tuned on emotion prediction

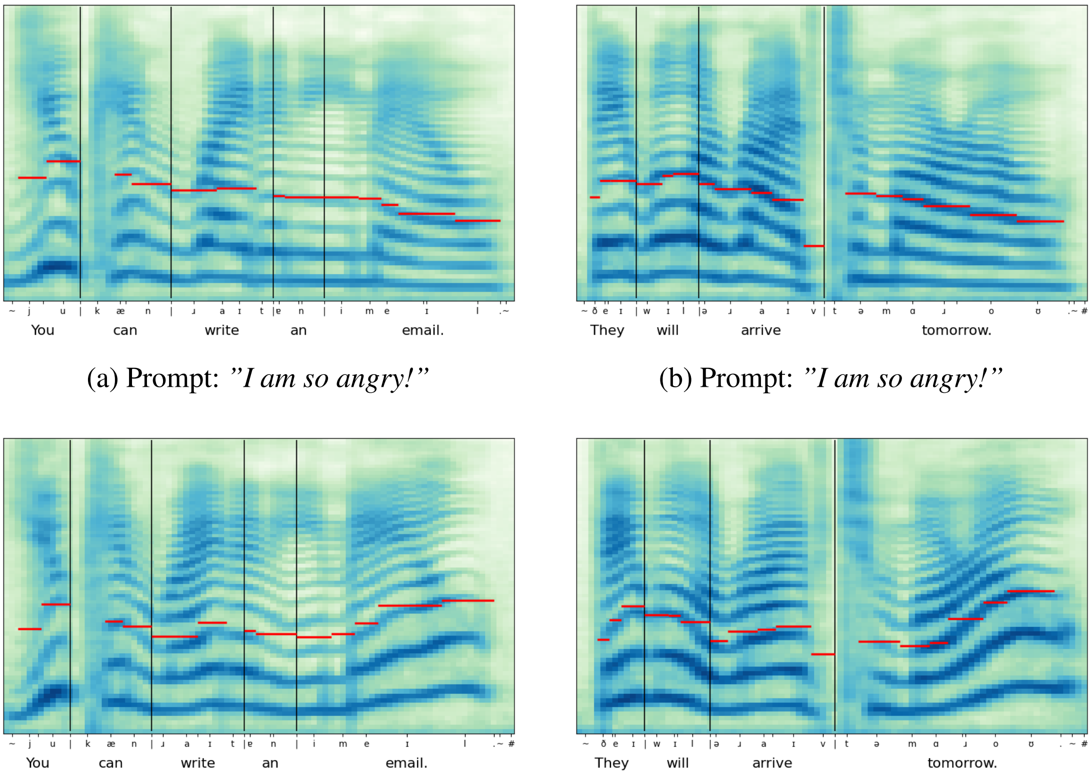


(a) Extracted with BERT

(b) Extracted with Sentence Transformer

Figure 30: T-SNE Plots of alternative Sentence Embeddings for Training Prompts

Figure 32 shows spectrograms for the same utterance generated with semantically similar prompts for the two emotion categories respectively. As can be seen, the produced spectrograms and pitch contours look almost identical which verifies that semantically similar prompts result in similar prosody.



(a) Prompt: *"I am so angry!"*

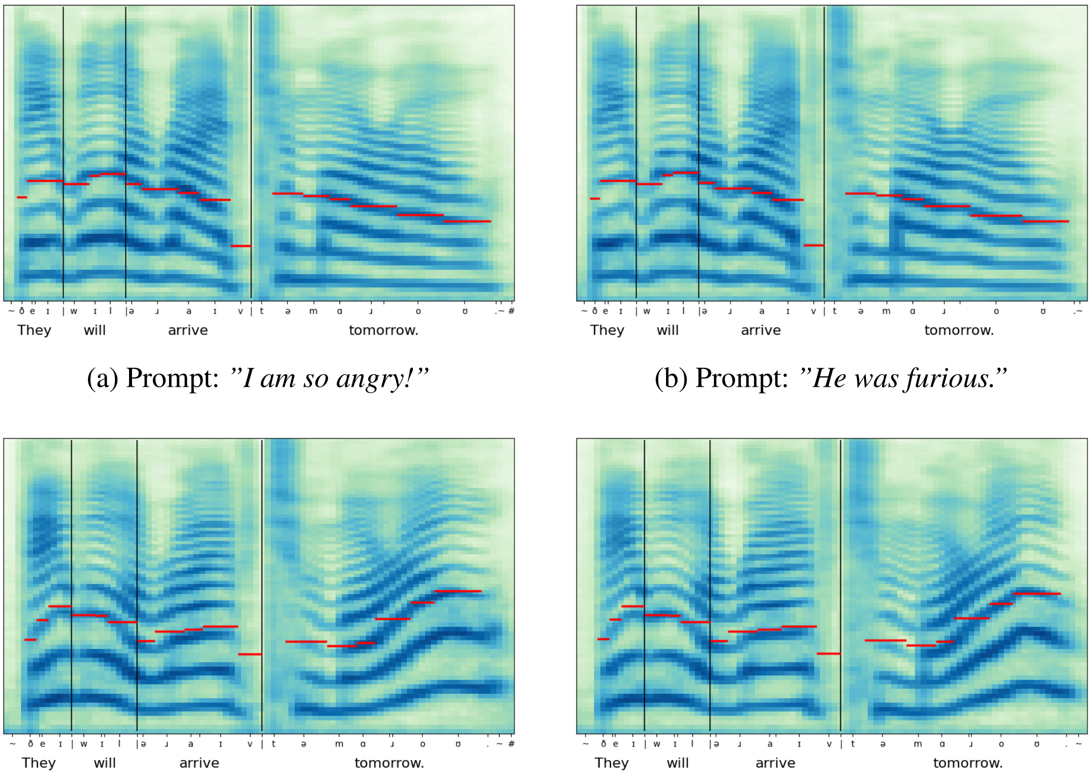
(b) Prompt: *"I am so angry!"*

(c) Prompt: *"What a surprise!"*

(d) Prompt: *"What a surprise!"*

Figure 31: Spectrograms generated with identical prompts

Horizontal lines in red indicate the pitch of each phoneme.



(a) Prompt: *"I am so angry!"*

(b) Prompt: *"He was furious."*

(c) Prompt: *"What a surprise!"*

(d) Prompt: *"She didn't expect that."*

Figure 32: Spectrograms generated with semantically similar prompts

Horizontal lines in red indicate the pitch of each phoneme.

8 Limitations and Future Work

While the TTS system proposed in this thesis allows to accurately obtain the desired emotional prosody through natural language prompts with key advantages over previous approaches, it also has several limitations.

As for now, it has been found in preliminary testing, that using pre-trained speaker embeddings during training instead of a look-up table does not work well. The main problem thereby is that such speaker embeddings seem to contain speaking style related information in addition to speaker timbre. This additionally influences the speech prosody, dampens the effect of the prompt or mixes it up completely. In future work it could be explored if advanced techniques for disentangling speaker timbre from prosodic information in speaker embeddings (e.g. An et al. (2022)) provide applicable solutions. This would be beneficial in the sense that it allows to synthesize utterances in the voice of speakers unseen during training in a zero-shot or few-shot manner.

Furthermore, the proposed model is not capable of performing cross-speaker emotion or style transfer since the prediction of prosodic parameters is highly dependent on the joint representation of specific speaker and prompt embeddings. However, there are various approaches which tackle this problem and could be adopted for the system in future work (e.g. Wu et al. (2021)).

Another limitation that arises is that there is no way of accounting for emotional intensity since this information is not present in the training data. It would be required to somehow obtain the intensity of both speech and textual prompt and match them accordingly such that the model can learn this correspondence.
Apart from that, it is not possible to model multiple emotions within a sentence, because the prosodic parameters are modeled over the whole utterance. However, this scenario might only appear when synthesizing long paragraphs which then could also be splitted up into smaller parts.

In future research the proposed TTS system could also be extended to multiple languages. The IMS Toucan Speech Synthesis Toolkit already provides the necessary components, however one would have to find suitable annotated multi-lingual expressive speech and text datasets. Multi-lingual language models which could be used for sentence embedding extraction are publicly available (Artetxe and Schwenk, 2019; Mao and Nakagawa, 2023), although, it might be required to fine-tune them on e.g. emotion prediction to make them useful for TTS training.

Finally, besides training the TTS system specifically on emotional speech, using other data such as audio-books could be explored similarly. Thereby, a method would have to be developed that obtains speech samples and textual prompts with high correspondence. Furthermore, context sentences of an utterance could be considered as candidate prompts.

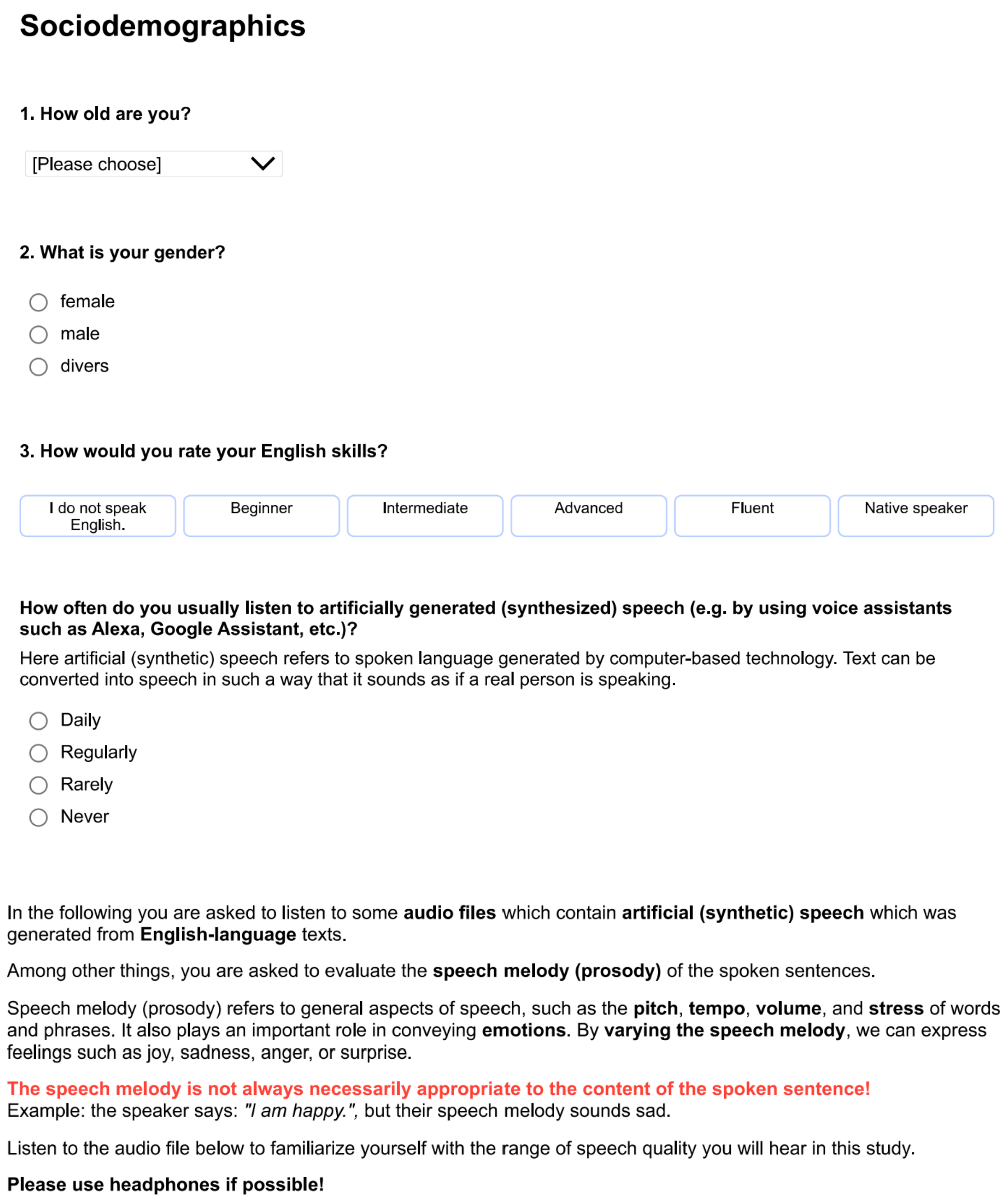
9 Conclusion

This thesis proposes a text-to-speech system that is conditioned on sentences embeddings extracted from natural language prompts in order to make the prosodic parameters of generated speech controllable in an intuitive and effective way. The system builds on a baseline which combines architectural designs from several TTS models proposed in previous work and provides benefits regarding speed, data efficiency, robustness and controllability. The proposed integration scheme concatenates speaker and sentence embeddings by modeling inter-dependencies between them before inducing the joint representation into the model's encoder, decoder and prosody predictors. This approach is simple and adds only little computational cost in comparison to the baseline. Furthermore, a training strategy is developed that operates on merged emotional speech and text datasets and varies prompts in each iteration, increasing the generalization capabilities of the model and reducing the risk of over-fitting. In addition, combining existing datasets eliminates the necessity of manually providing style descriptions along with the speech data. Extensive objective and subjective evaluations on utterances generated from sentences of emotional text datasets demonstrate the prompting capabilities of the conditioned TTS system. It outperforms the baseline with regard to the degree of prosodic controllability as indicated by emotion classification and A/B preference tests. Thereby, the emotional content of the provided prompt is transferred accurately to the prosody of generated speech. Moreover, prosodic similarity tests suggest that the produced prosody is solely inferred from the prompt and not influenced by the input text. At the same time the system maintains precise controllability of speaker identities, as indicated by cosine similarity scores, and overall high speech quality and intelligibility, as suggested by collected mean opinion scores and calculated word error rates. Besides a high correlation between prompts and speech prosody in the training data, fine-tuning the language model, which is used for sentence embedding extraction, has been found to be crucial. In this thesis a distilled RoBERTa model, fine-tuned on the task of emotion prediction, has been applied and its adequateness compared to e.g. standard BERT embeddings has been shown in T-SNE plots.

The proposed TTS system has several limitations which could be addressed in future research. Currently, the model is not capable of synthesizing utterances in the voice of speaker unseen during training or performing cross-speaker style transfer. Furthermore, it cannot account for emotional intensities or mixed emotions. Finally, the system could be extended for multi-lingual purposes.

10 Appendix

10.1 Questions of the Subjective Listening Test



Sociodemographics

1. How old are you?

[Please choose] ▾

2. What is your gender?

○ female
○ male
○ divers

3. How would you rate your English skills?

I do not speak English.	Beginner	Intermediate	Advanced	Fluent	Native speaker

How often do you usually listen to artificially generated (synthesized) speech (e.g. by using voice assistants such as Alexa, Google Assistant, etc.)?

Here artificial (synthetic) speech refers to spoken language generated by computer-based technology. Text can be converted into speech in such a way that it sounds as if a real person is speaking.

○ Daily
○ Regularly
○ Rarely
○ Never

In the following you are asked to listen to some **audio files** which contain **artificial (synthetic) speech** which was generated from **English-language** texts.

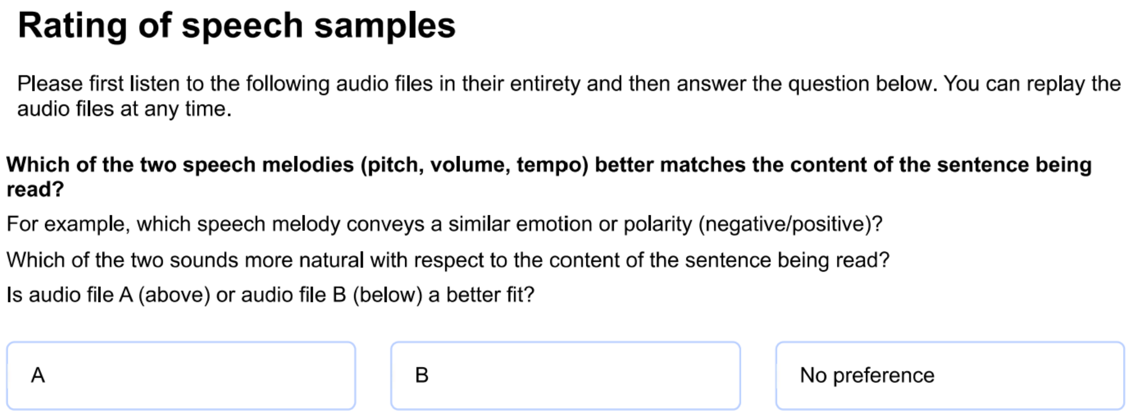
Among other things, you are asked to evaluate the **speech melody (prosody)** of the spoken sentences.

Speech melody (prosody) refers to general aspects of speech, such as the **pitch**, **tempo**, **volume**, and **stress** of words and phrases. It also plays an important role in conveying **emotions**. By **varying the speech melody**, we can express feelings such as joy, sadness, anger, or surprise.

The speech melody is not always necessarily appropriate to the content of the spoken sentence!
Example: the speaker says: *"I am happy.",* but their speech melody sounds sad.

Listen to the audio file below to familiarize yourself with the range of speech quality you will hear in this study.

Please use headphones if possible!



Rating of speech samples

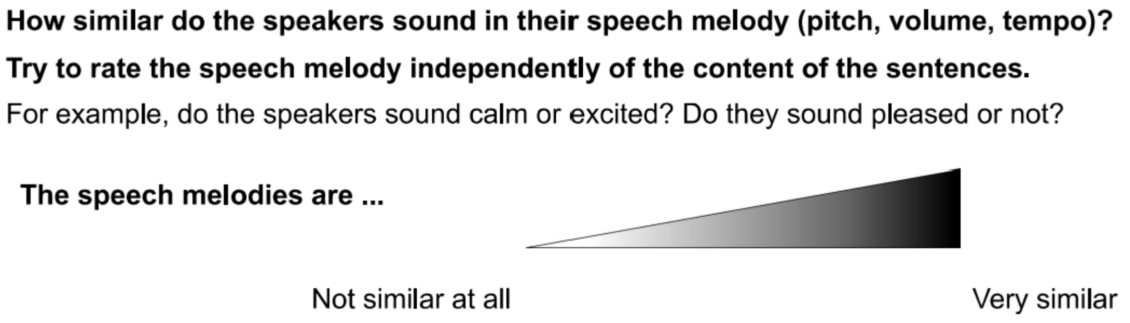
Please first listen to the following audio files in their entirety and then answer the question below. You can replay the audio files at any time.

Which of the two speech melodies (pitch, volume, tempo) better matches the content of the sentence being read?

For example, which speech melody conveys a similar emotion or polarity (negative/positive)?

Which of the two sounds more natural with respect to the content of the sentence being read?

Is audio file A (above) or audio file B (below) a better fit?

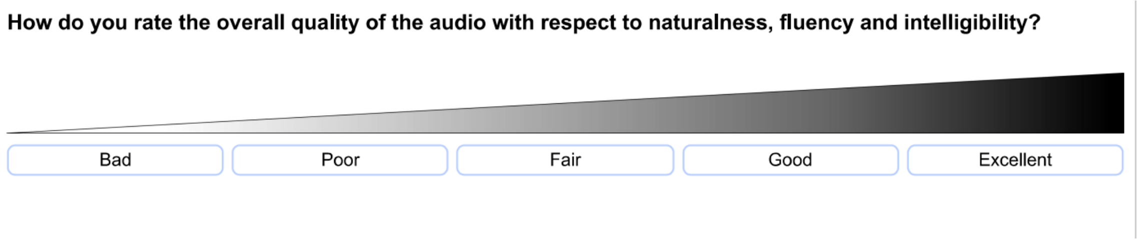


How similar do the speakers sound in their speech melody (pitch, volume, tempo)?

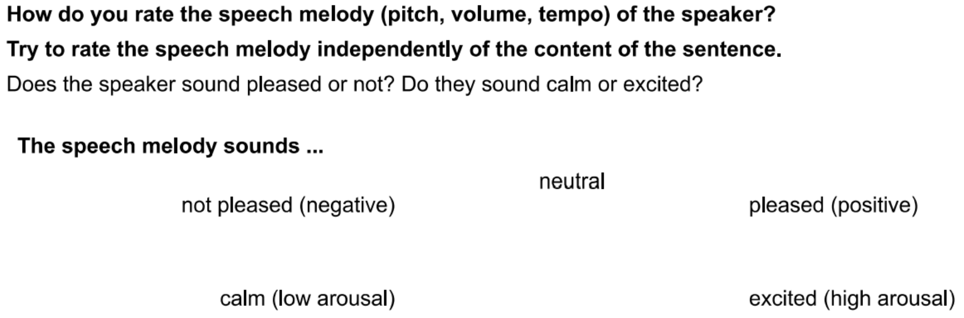
Try to rate the speech melody independently of the content of the sentences.

For example, do the speakers sound calm or excited? Do they sound pleased or not?

The speech melodies are ...



How do you rate the overall quality of the audio with respect to naturalness, fluency and intelligibility?

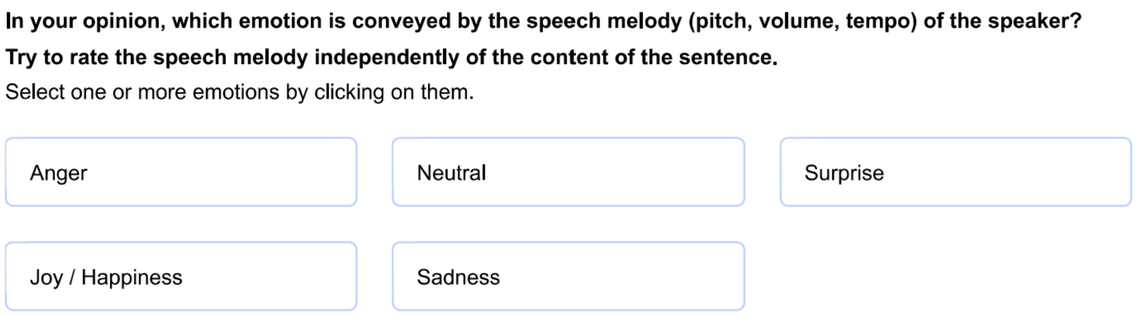


How do you rate the speech melody (pitch, volume, tempo) of the speaker?

Try to rate the speech melody independently of the content of the sentence.

Does the speaker sound pleased or not? Do they sound calm or excited?

The speech melody sounds ...

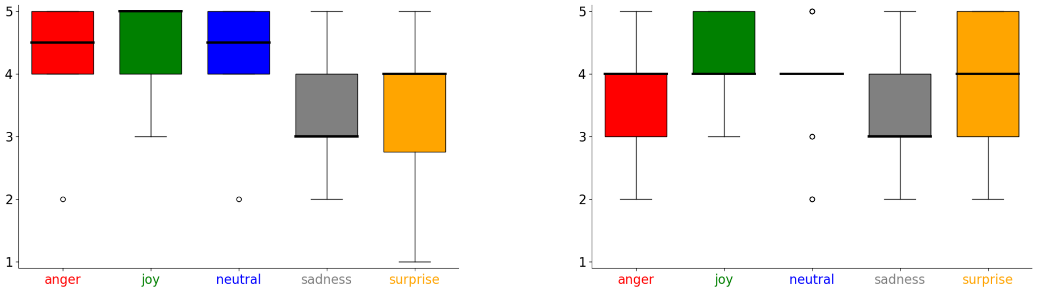


In your opinion, which emotion is conveyed by the speech melody (pitch, volume, tempo) of the speaker?

Try to rate the speech melody independently of the content of the sentence.

Select one or more emotions by clicking on them.

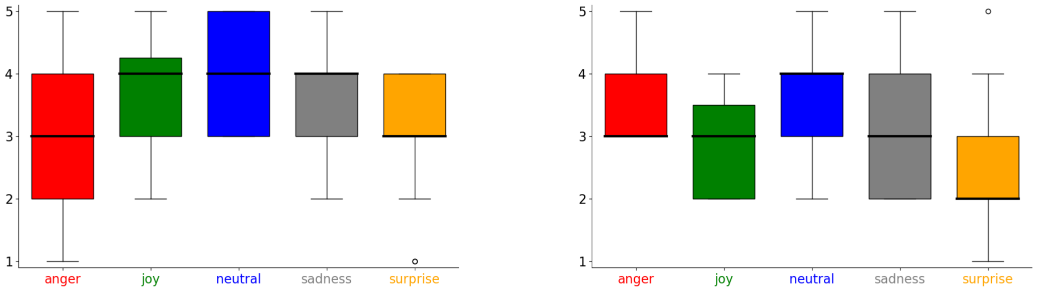
10.2 Box-Plots Mean Opinion Score



(a) female

(b) male

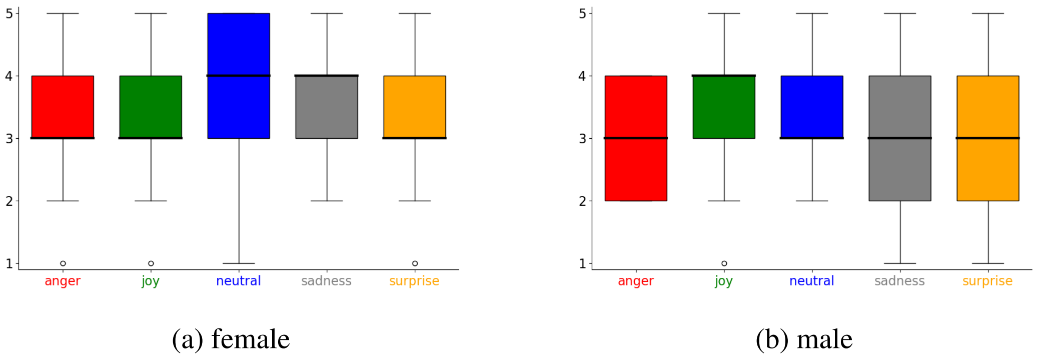
Figure 33: Mean Opinion Scores, Ground Truth



(a) female

(b) male

Figure 34: Mean Opinion Scores, Baseline



(a) female

(b) male

Figure 35: Mean Opinion Scores, Proposed

10.3 Valence-Arousal Ratings

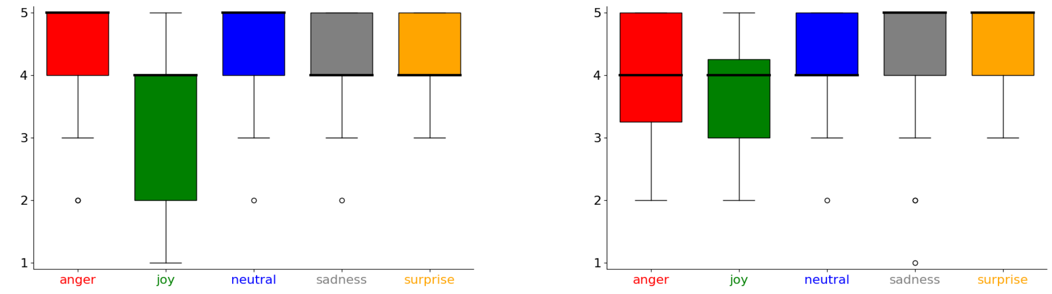
		anger	**joy**	**neutral**	**sadness**	**surprise**
Ground Truth	female	2.31	4.06	3.53	2.88	3.50
	male	3.40	4.56	3.58	2.62	3.58
Baseline	female	2.53	2.88	3.00	2.13	2.71
	male	2.75	2.40	3.25	2.41	2.29
Proposed Same	female	1.44	3.31	3.00	1.88	3.25
	male	2.31	4.00	2.56	2.29	3.76
Proposed Other	female	1.65	3.00	2.06	2.59	3.67
	male	3.00	2.67	2.40	2.29	2.47

Table 9: Valence Ratings

		anger	**joy**	**neutral**	**sadness**	**surprise**
Ground Truth	female	5.00	3.50	2.53	2.00	4.25
	male	3.44	4.00	2.81	1.67	4.59
Baseline	female	3.19	2.31	3.00	3.06	2.47
	male	3.50	2.25	3.50	2.18	1.71
Proposed Same	female	4.40	2.63	2.12	1.80	4.00
	male	2.40	3.25	2.25	2.29	3.44
Proposed Other	female	4.53	3.19	2.47	2.40	4.21
	male	3.38	3.50	1.94	1.94	4.00

Table 10: Arousal Ratings

10.3.1 Box-Plots Prosodic Similarity

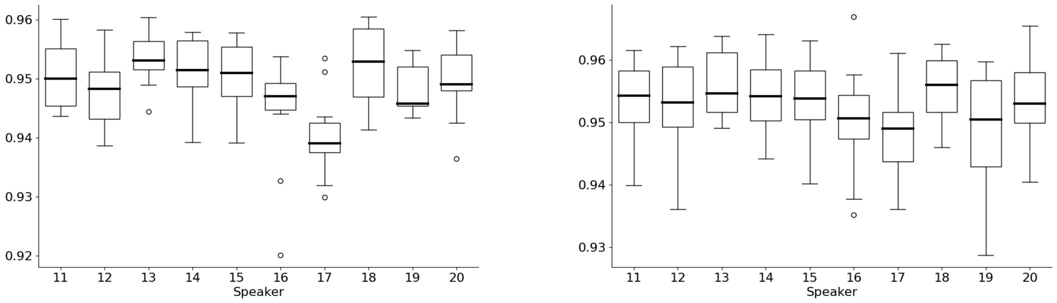


(a) Female

(b) Male

Figure 36: Prosodic Similarity Ratings

10.4 Box-Plots Speaker Similarity

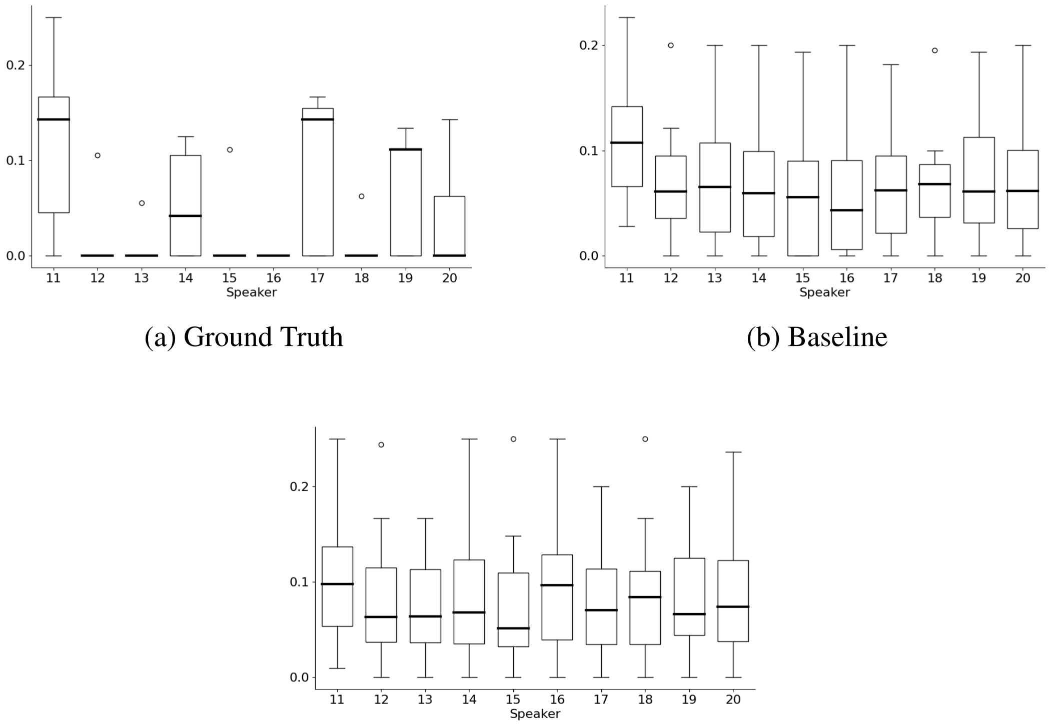


(a) Baseline

(b) Proposed

Figure 37: Speaker Similarities

10.5 Box-Plots Word Error Rate



(a) Ground Truth

(b) Baseline

(c) Proposed

Figure 38: Word Error Rates

11 References

Cecilia Ovesdotter Alm and Richard Sproat. Perceptions of emotions in expressive storytelling. In *Ninth European Conference on Speech Communication and Technology*, 2005.

Cecilia Ovesdotter Alm, Dan Roth, and Richard Sproat. Emotions from text: Machine learning for text-based emotion prediction. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 579–586, Vancouver, British Columbia, Canada, October 2005. Association for Computational Linguistics. URL `https://aclanthology.org/H05-1073`.

Xiaochun An, Frank K. Soong, and Lei Xie. Disentangling style and speaker attributes for TTS style transfer. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:646–658, 2022. doi: 10.1109/taslp.2022.3145297.

Mikel Artetxe and Holger Schwenk. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610, nov 2019. doi: 10.1162/tacl_a_00288.

Nabiha Asghar. Yelp dataset challenge: Review rating prediction. *arXiv preprint arXiv:1605.05362*, 2016.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Jo-Anne Bachorowski. Vocal expression and perception of emotion. *Current Directions in Psychological Science*, 8(2):53–57, apr 1999. doi: 10.1111/1467-8721.00013.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460. Curran Associates, Inc., 2020. URL `https://proceedings.neurips.cc/paper_files/paper/2020/file/92d1e1eb1cd6f9fba3227870bb6d7f07-Paper.pdf`.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

Taejun Bak, Jae-Sung Bae, Hanbin Bae, Young-Ik Kim, and Hoon-Young Cho. Fastpitchformant: Source-filter based decomposed modeling for speech synthesis. *Conference of the International Speech CommunicationAssociation (Interspeech), 2021*, 2021.

Taejun Bak, Junmo Lee, Hanbin Bae, Jinhyeok Yang, Jae-Sung Bae, and Young-Sun Joo. Avocodo: Generative adversarial network for artifact-free vocoder. *arXiv preprint arXiv:2206.13404*, 2022.

Philip Bard. A diencephalic mechanism for the expression of rage with special reference to the sympathetic nervous system. *American Journal of Physiology-Legacy Content*, 84(3):490–515, 1928.

Lisa Feldman Barrett. *How emotions are made: The secret life of the brain*. Pan Macmillan, 2017.

Stefan Baumann and Arndt Riester. Referential and lexical givenness: Semantic, prosodic and cognitive aspects. *Prosody and meaning*, 25:119–162, 2012.

Yoshua Bengio, Réjean Ducharme, and Pascal Vincent. A neural probabilistic language model. *Advances in neural information processing systems*, 13, 2003.

Patricia EG Bestelmeyer, Sonja A Kotz, and Pascal Belin. Effects of emotional valence and arousal on the voice perception network. *Social cognitive and affective neuroscience*, 12 (8):1351–1358, 2017.

James Betker. Better speech synthesis through scaling. *arXiv preprint arXiv:2305.07243*, 2023.

BigScience-Workshop. Bloom: A 176b-parameter open-access multilingual language model, 2022. URL `https://arxiv.org/abs/2211.05100`.

Alan W Black, Heiga Zen, and Keiichi Tokuda. Statistical parametric speech synthesis. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, volume 4, pages IV–1229. IEEE, 2007.

Antoine Bordes, Sumit Chopra, and Jason Weston. Question answering with subgraph embeddings. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2014. doi: 10.3115/v1/d14-1067.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42: 335–359, 2008.

Walter B Cannon. The james-lange theory of emotions: A critical examination and an alternative theory. *The American journal of psychology*, 39(1/4):106–124, 1927.

Liping Chen, Yan Deng, Xi Wang, Frank K. Soong, and Lei He. Speech bert embedding for improving prosody in neural TTS. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, jun 2021. doi: 10.1109/icassp39728.2021.9413864.

Mingjian Chen, Xu Tan, Yi Ren, Jin Xu, Hao Sun, Sheng Zhao, and Tao Qin. MultiSpeech: Multi-speaker text to speech with transformer. In *Interspeech 2020*. ISCA, oct 2020. doi: 10.21437/interspeech.2020-3139.

S.F. Chen and R. Rosenfeld. A survey of smoothing techniques for ME models. *IEEE Transactions on Speech and Audio Processing*, 8(1):37–50, 2000. doi: 10.1109/89.817452.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2014. doi: 10.3115/v1/d14-1179.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. Natural language processing (almost) from scratch. *Journal of machine learning research*, 12(ARTICLE):2493–2537, 2011.

Jian Cong, Shan Yang, Na Hu, Guangzhi Li, Lei Xie, and Dan Su. Controllable context-aware conversational speech synthesis. In *Interspeech 2021*. ISCA, aug 2021. doi: 10.21437/interspeech.2021-412.

Tobias Cornille, Fengna Wang, and Jessa Bekker. Interactive multi-level prosody control for expressive speech synthesis. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, may 2022. doi: 10.1109/icassp43922.2022.9746654.

Harald Cramér. *Mathematical methods of statistics*, volume 43. Princeton university press, 1999.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc Le, and Ruslan Salakhutdinov. Transformer-XL: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2978–2988, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1285. URL `https://aclanthology.org/P19-1285`.

Charles Darwin. *The expression of the emotions in man and animals*. Oxford University Press, USA, 1872.

Yann N Dauphin, Angela Fan, Michael Auli, and David Grangier. Language modeling with gated convolutional networks. In *International conference on machine learning*, pages 933–941. PMLR, 2017.

Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification. In *Interspeech 2020*. ISCA, oct 2020. doi: 10.21437/interspeech.2020-2650.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers forlanguage understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*, pages 4171–4186. Association for Computational Linguistics, June 2019. doi: 10.18653/v1/n19-1423.

Chenpeng Du and Kai Yu. Phone-level prosody modelling with GMM-based MDN for diverse and controllable speech synthesis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:190–201, 2022. doi: 10.1109/taslp.2021.3133205.

Paul Ekman et al. Basic emotions. *Handbook of cognition and emotion*, 98(45-60):16, 1999.

Wei Fang, Yu-An Chung, and James Glass. Towards transfer learning for end-to-end speech synthesis from deep pre-trained language models. June 2019.

Angela D Friederici. Syntactic, prosodic, and semantic processes in the brain: evidence from event-related neuroimaging. *Journal of Psycholinguistic Research*, 30:237–250, 2001.

Aaron Gokaslan and Vanya Cohen. Openwebtext corpus. `http://Skylion007.github.io/OpenWebTextCorpus`, 2019.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. URL `https://proceedings.neurips.cc/paper_files/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf`.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. Connectionist temporal classification. In *Proceedings of the 23rd international conference on Machine learning - ICML '06*. ACM Press, 2006. doi: 10.1145/1143844.1143891.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al. Conformer: Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv:2005.08100*, 2020.

Haohan Guo, Frank K. Soong, Lei He, and Lei Xie. Exploiting syntactic features in a parsed tree to improve end-to-end TTS. In *Interspeech 2019*. ISCA, sep 2019. doi: 10.21437/interspeech.2019-2167.

Pengcheng Guo, Florian Boyer, Xuankai Chang, Tomoki Hayashi, Yosuke Higuchi, Hirofumi Inaguma, Naoyuki Kamo, Chenda Li, Daniel Garcia-Romero, Jiatong Shi, Jing Shi, Shinji Watanabe, Kun Wei, Wangyou Zhang, and Yuekai Zhang. Recent developments on espnet toolkit boosted by conformer. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, jun 2021. doi: 10.1109/icassp39728.2021.9414858.

Yiwei Guo, Chenpeng Du, and Kai Yu. Unsupervised word-level prosody tagging for controllable speech synthesis. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7597–7601. IEEE, 2022a.

Zhifang Guo, Yichong Leng, Yihan Wu, Sheng Zhao, and Xu Tan. Prompttts: Controllable text-to-speech with text descriptions. 2022b. doi: 10.48550/ARXIV.2211.12171. URL `https://arxiv.org/abs/2211.12171`.

Zellig S Harris. Distributional structure. *Word*, 10(2-3):146–162, 1954.

Jochen Hartmann. Emotion english distilroberta-base. `https://huggingface.co/j-hartmann/emotion-english-distilroberta-base/`, 2022.

Tomoki Hayashi, Shinji Watanabe, Tomoki Toda, Kazuya Takeda, Shubham Toshniwal, and Karen Livescu. Pre-trained text embeddings for enhanced text-to-speech synthesis. In *Interspeech 2019*. ISCA, sep 2019. doi: 10.21437/interspeech.2019-3177.

Geoffrey Hinton, Li Deng, Dong Yu, George Dahl, Abdel rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara Sainath, and Brian Kingsbury. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6):82–97, nov 2012. doi: 10.1109/msp.2012.2205597.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Geoffrey E Hinton and Sam Roweis. Stochastic neighbor embedding. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems*, volume 15. MIT Press, 2002. URL `https://proceedings.neurips.cc/paper_files/paper/2002/file/6150ccc6069bea6b5716254057a194ef-Paper.pdf`.

Julia Hirschberg and Owen Rambow. Learning prosodic features using a tree representation. In *Seventh European Conference on Speech Communication and Technology*, 2001.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, nov 1997. doi: 10.1162/neco.1997.9.8.1735.

John J Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences*, 79(8): 2554–2558, 1982.

Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

Andrew J Hunt and Alan W Black. Unit selection in a concatenative speech synthesis system using a large speech database. In *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, volume 1, pages 373–376. IEEE, 1996.

International-Phonetic-Association. *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*. Cambridge University Press, 1999.

Keith Ito and Linda Johnson. The lj speech dataset. `https://keithito.com/LJ-Speech-Dataset/`, 2017.

William James. What is emotion? 1884. 1884.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association* *for Computational Linguistics*. Association for Computational Linguistics, 2019. doi: 10.18653/v1/p19-1356.

Ye Jia, Yu Zhang, Ron Weiss, Quan Wang, Jonathan Shen, Fei Ren, Patrick Nguyen, Ruoming Pang, Ignacio Lopez Moreno, Yonghui Wu, et al. Transfer learning from speaker verification to multispeaker text-to-speech synthesis. *Advances in neural information processing systems*, 31, 2018.

Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438, dec 2020. doi: 10.1162/tacl_a_00324.

Sri Karlapati, Ammar Abbas, Zack Hodari, Alexis Moinet, Arnaud Joly, Penny Karanasou, and Thomas Drugman. Prosodic representation learning and contextual sampling for neural text-to-speech. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, jun 2021. doi: 10.1109/icassp39728.2021.9413696.

Tom Kenter, Manish Sharma, and Rob Clark. Improving the prosody of RNN-based english text-to-speech synthesis by incorporating a BERT model. In *Interspeech 2020*. ISCA, oct 2020. doi: 10.21437/interspeech.2020-1430.

Jaehyeon Kim, Sungwon Kim, Jungil Kong, and Sungroh Yoon. Glow-tts: A generative flow for text-to-speech via monotonic alignment search. *Advances in Neural Information Processing Systems*, 33:8067–8077, 2020.

Jaehyeon Kim, Jungil Kong, and Juhee Son. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In *International Conference on Machine Learning*, pages 5530–5540. PMLR, 2021a.

Minchan Kim, Sung Jun Cheon, Byoung Jin Choi, Jong Jin Kim, and Nam Soo Kim. Expressive text-to-speech using style tag. In *Interspeech 2021*. ISCA, aug 2021b. doi: 10.21437/interspeech.2021-465.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Julia Koch, Florian Lux, Nadja Schauffler, Toni Bernhart, Felix Dieterle, Jonas Kuhn, Sandra Richter, Gabriel Viehhauser, and Ngoc Thang Vu. PoeticTTS - controllable poetry reading for literary studies. In *Interspeech 2022*. ISCA, sep 2022. doi: 10.21437/interspeech.2022-10841.

Yuma Koizumi, Heiga Zen, Shigeki Karita, Yifan Ding, Kohei Yatabe, Nobuyuki Morioka, Michiel Bacchiani, Yu Zhang, Wei Han, and Ankur Bapna. Libritts-r: A restored multi-speaker text-to-speech corpus. *arXiv preprint arXiv:2305.18802*, 2023a.

Yuma Koizumi, Heiga Zen, Shigeki Karita, Yifan Ding, Kohei Yatabe, Nobuyuki Morioka, Yu Zhang, Wei Han, Ankur Bapna, and Michiel Bacchiani. Miipher: A robust speech restoration model integrating self-supervised speech and text representations. *arXiv preprint arXiv:2303.01664*, 2023b.

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 17022–17033. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/c5d736809766d46260d816d8dbc9eb44-Paper.pdf.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. URL https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

Kundan Kumar, Rithesh Kumar, Thibault de Boissiere, Lucas Gestin, Wei Zhen Teoh, Jose Sotelo, Alexandre de Brébisson, Yoshua Bengio, and Aaron C Courville. Melgan: Generative adversarial networks for conditional waveform synthesis. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/6804c9bca0a615bdb9374d00a9fcba59-Paper.pdf.

Arne Köhn, Timo Baumann, and Oskar Dörfler. An empirical analysis of the correlation of syntax and prosody. In *Interspeech 2018*. ISCA, sep 2018. doi: 10.21437/interspeech.2018-2530.

Adrian Lancucki. Fastpitch: Parallel text-to-speech with pitch prediction. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, jun 2021. doi: 10.1109/icassp39728.2021.9413889.

Carl Georg Lange and Hans Kurella. *Ueber Gemüthsbewegungen: eine psycho-physiologische Studie*. T. Thomas, 1885.

Sang-gil Lee, Wei Ping, Boris Ginsburg, Bryan Catanzaro, and Sungroh Yoon. Bigvgan: A universal neural vocoder with large-scale training. *arXiv preprint arXiv:2206.04658*, 2022.

A. F G Leentjens, S. M Wielaert, F. van Harskamp, and F. W Wilmink. Disturbances of affective prosody in patients with schizophrenia, a cross sectional study. *J Neurol Neurosurg Psychiatry*, 64(3):375–378, mar 1998. doi: 10.1136/jnnp.64.3.375.

Naihan Li, Shujie Liu, Yanqing Liu, Sheng Zhao, and Ming Liu. Neural speech synthesis with transformer network. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33:6706–6713, jul 2019. doi: 10.1609/aaai.v33i01.33016706.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan, November 2017. Asian Federation of Natural Language Processing. URL `https://aclanthology.org/I17-1099`.

Yinghao Aaron Li, Cong Han, and Nima Mesgarani. Styletts: A style-based generative model for natural and diverse text-to-speech synthesis. *arXiv preprint arXiv:2205.15439*, 2022.

Guanghou Liu, Yongmao Zhang, Yi Lei, Yunlin Chen, Rui Wang, Zhifei Li, and Lei Xie. Promptstyle: Controllable style transfer for text-to-speech with natural language descriptions. *arXiv preprint arXiv:2305.19522*, 2023.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, sep 2022. doi: 10.1145/3560815.

Rui Liu, Berrak Sisman, and Haizhou Li. Graphspeech: Syntax-aware graph attention network for neural speech synthesis. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6059–6063, Toronto, ON, Canada, 2021a. IEEE. ISBN 978-1-7281-7606-2. doi: 10.1109/ICASSP39728.2021.9413513.

Yanqing Liu, Zhihang Xu, Gang Wang, Kuan Chen, Bohan Li, Xu Tan, Jinzhu Li, Lei He, and Sheng Zhao. Delightfultts: The microsoft speech synthesis system for blizzard challenge 2021. *arXiv preprint arXiv:2110.12612*, 2021b.

Steven R Livingstone and Frank A Russo. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PloS one*, 13(5):e0196391, 2018.

Yiping Lu, Zhuohan Li, Di He, Zhiqing Sun, Bin Dong, Tao Qin, Liwei Wang, and Tie-Yan Liu. Understanding and improving transformer from a multi-particle dynamic system point of view. *arXiv preprint arXiv:1906.02762*, 2019.

Florian Lux and Thang Vu. Language-agnostic meta-learning for low-resource text-to-speech with articulatory features. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.acl-long.472.

Florian Lux, Julia Koch, Antje Schweitzer, and Thang Vu. The ims toucan system for the blizzard challenge 2021. In *Blizzard Challenge 2021*, 2021.

Florian Lux, Julia Koch, and Ngoc Thang Vu. Low-resource multilingual and zero-shot multispeaker TTS. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 741–751, Online only, November 2022. Association for Computational Linguistics. URL `https://aclanthology.org/2022.aacl-main.56`.

Florian Lux, Julia Koch, Sarina Meyer, Thomas Bott, Nadja Schauffler, Pavel Denisov, Antje Schweitzer, and Ngoc Thang Vu. The IMS Toucan system for the Blizzard Challenge 2023. In *Proc. Blizzard Challenge Workshop*. Speech Synthesis SIG, 2023a.

Florian Lux, Julia Koch, and Ngoc Thang Vu. Exact prosody cloning in zero-shot multispeaker text-to-speech. In *2022 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, jan 2023b. doi: 10.1109/slt54892.2023.10022433.

Junshui Ma, Robert P. Sheridan, Andy Liaw, George E. Dahl, and Vladimir Svetnik. Deep neural nets as a method for quantitative structure–activity relationships. *Journal of Chemical Information and Modeling*, 55(2):263–274, feb 2015. doi: 10.1021/ci500747n.

Zhuoyuan Mao and Tetsuji Nakagawa. Lealla: Learning lightweight language-agnostic sentence embeddings with knowledge distillation, 2023. URL `https://arxiv.org/abs/2302.08387`.

Warren S McCulloch and Walter Pitts. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5:115–133, 1943.

Brian McFee, Colin Raffel, Dawen Liang, Daniel P Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference*, volume 8, pages 18–25, 2015.

Sarina Meyer, Florian Lux, Julia Koch, Pavel Denisov, Pascal Tilli, and Ngoc Thang Vu. Prosody is not identity: A speaker anonymization approach using prosody cloning. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, jun 2023. doi: 10.1109/icassp49357.2023.10096607.

Chenfeng Miao, Shuang Liang, Minchuan Chen, Jun Ma, Shaojun Wang, and Jing Xiao. Flow-TTS: A non-autoregressive network for text to speech based on flow. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, may 2020. doi: 10.1109/icassp40776.2020.9054484.

Tomas Mikolov, Martin Karafiát, Lukas Burget, Jan Cernockỳ, and Sanjeev Khudanpur. Recurrent neural network based language model. In *Interspeech*, volume 2, pages 1045–1048. Makuhari, 2010.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

Huaiping Ming, Lei He, Haohan Guo, and Frank K. Soong. Feature reinforcement with word embedding and parsing information in neural tts. January 2019.

Taniya Mishra, Yeon jun Kim, and Srinivas Bangalore. Intonational phrase break prediction for text-to-speech synthesis using dependency relations. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, apr 2015. doi: 10.1109/icassp.2015.7178906.

Eric Moulines and Francis Charpentier. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech communication*, 9 (5-6):453–467, 1990.

Arijit Mukherjee, Shubham Bansal, Sandeepkumar Satpal, and Rupesh Mehta. Text aware emotional text-to-speech with BERT. In *Interspeech 2022*. ISCA, sep 2022. doi: 10.21437/interspeech.2022-11293.

Shifeng Pan and Lei He. Cross-speaker style transfer with prosody bottleneck in neural speech synthesis. *arXiv preprint arXiv:2107.12562*, 2021.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: An ASR corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, apr 2015. doi: 10.1109/icassp.2015.7178964.

Marc D. Pell and Sonja A. Kotz. On the time course of vocal emotion recognition. *PLoS ONE*, 6(11):e27256, nov 2011. doi: 10.1371/journal.pone.0027256.

Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2014. doi: 10.3115/v1/d14-1162.

Alejandro Pérez-González-de Martos, Albert Sanchis, and Alfons Juan. Vrain-upv mllp's system for the blizzard challenge 2021. *Proc. Blizzard ChallengeWorkshop*, 2021.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, 2019. doi: 10.18653/v1/d19-1250.

M. Kathleen Pichora-Fuller and Kate Dupuis. Toronto emotional speech set (tess), 2020.

Wei Ping, Kainan Peng, Andrew Gibiansky, Sercan Ömer Arik, Ajay Kannan, Sharan Narang, Jonathan Raiman, and John Miller. Deep voice 3: 2000-speaker neural text-to-speech. *CoRR*, abs/1710.07654, 2017. URL `http://arxiv.org/abs/1710.07654`.

Robert Plutchik. A general psychoevolutionary theory of emotion. In *Theories of emotion*, pages 3–33. Elsevier, 1980.

Vadim Popov, Ivan Vovk, Vladimir Gogoryan, Tasnima Sadekova, and Mikhail Kudinov. Grad-tts: A diffusion probabilistic model for text-to-speech. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8599–8608. PMLR, 18–24 Jul 2021. URL `https://proceedings.mlr.press/v139/popov21a.html`.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1), jan 2020. ISSN 1532-4435.

Tuomo Raitio, Ramya Rasipuram, and Dan Castellani. Controllable neural text-to-speech synthesis using intuitive prosodic features. In *Interspeech 2020*. ISCA, oct 2020. doi: 10.21437/interspeech.2020-2861.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, Abdelwahab Heba, Jianyuan Zhong, Ju-Chieh Chou, Sung-Lin Yeh, Szu-Wei Fu, Chien-Feng Liao, Elena Rastorgueva, François Grondin, William Aris, Hwidong Na, Yan Gao, Renato De Mori, and Yoshua Bengio. SpeechBrain: A general-purpose speech toolkit, 2021. arXiv:2106.04624.

Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *International conference on machine learning*, pages 1060–1069. PMLR, 2016.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, 2019. doi: 10.18653/v1/d19-1410.

Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. Fastspeech: Fast, robust and controllable text to speech. *33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada*, May 2019.

Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. Fastspeech 2: Fast and high-quality end-to-end text to speech. June 2020.

Yi Ren, Jinglin Liu, and Zhou Zhao. Portaspeech: Portable and high-quality generative text-to-speech. *35th Conference on Neural Information Processing Systems (NeurIPS 2021)*, September 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

Frank Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958.

James A Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161, 1980.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108, 2019.

Disa A. Sauter, Frank Eisner, Andrew J. Calder, and Sophie K. Scott. Perceptual cues in nonverbal vocal expressions of emotion. *Quarterly Journal of Experimental Psychology*, 63(11):2251–2272, nov 2010. doi: 10.1080/17470211003721642.

Ervin Sejdić, Igor Djurović, and Jin Jiang. Time–frequency feature representation using energy concentration: An overview of recent advances. *Digital Signal Processing*, 19 (1):153–183, jan 2009. doi: 10.1016/j.dsp.2007.12.004.

Amrith Setlur, Aman Madaan, Tanmay Parekh, Yiming Yang, and Alan W Black. Towards using heterogeneous relation graphs for end-to-end tts. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1162–1169, Cartagena, Colombia, 2021. IEEE. ISBN 978-1-6654-3740-0. doi: 10.1109/ASRU51503.2021.9687876.

Slava Shechtman and Alex Sorin. Sequence to sequence neural speech synthesis with prosody modification capabilities. In *10th ISCA Workshop on Speech Synthesis (SSW 10)*. ISCA, sep 2019. doi: 10.21437/ssw.2019-49.

Feiyu Shen, Chenpeng Du, and Kai Yu. Acoustic word embeddings for end-to-end speech synthesis. *Applied Sciences*, 11(19), 2021. ISSN 2076-3417. doi: 10.3390/app11199010. URL `https://www.mdpi.com/2076-3417/11/19/9010`.

Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, Rif A. Saurous, Yannis Agiomvrgiannakis, and Yonghui Wu. Natural TTS synthesis by conditioning wavenet on MEL spectrogram predictions. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, apr 2018. doi: 10.1109/icassp.2018.8461368.

Yookyung Shin, Younggun Lee, Suhee Jo, Yeongtae Hwang, and Taesu Kim. Text-driven emotional style control and cross-speaker style transfer in neural TTS. In *Interspeech 2022*. ISCA, sep 2022. doi: 10.21437/interspeech.2022-10131.

David Snyder, Daniel Garcia-Romero, Alan McCree, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur. Spoken language recognition using x-vectors. In *Odyssey*, volume 2018, pages 105–111, 2018.

Marlene Staib, Tian Huey Teh, Alexandra Torresquintero, Devang S. Ram Mohan, Lorenzo Foglianti, Raphael Lenain, and Jiameng Gao. Phonological features for 0-shot multilingual speech synthesis. In *Interspeech 2020*. ISCA, oct 2020. doi: 10.21437/interspeech.2020-1821. URL `https://doi.org/10.21437%2Finterspeech.2020-1821`.

Christian J Steinmetz and Joshua Reiss. pyloudnorm: A simple yet flexible loudness meter in python. In *Audio Engineering Society Convention 150*. Audio Engineering Society, 2021.

Stanley Smith Stevens, John Volkmann, and Edwin Broomell Newman. A scale for the measurement of the psychological magnitude pitch. *The journal of the acoustical society of america*, 8(3):185–190, 1937.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. URL `https://proceedings.neurips.cc/paper_files/paper/2014/file/a14ac55a4f27472c5d894ec1c3c743d2-Paper.pdf`.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2019. doi: 10.18653/v1/p19-1452.

Keiichi Tokuda, Takayoshi Yoshimura, Takashi Masuko, Takao Kobayashi, and Tadashi Kitamura. Speech parameter generation algorithms for hmm-based speech synthesis. In *2000 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 00CH37100)*, volume 3, pages 1315–1318. IEEE, 2000.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Jianhong Tu, Zeyu Cui, Xiaohuan Zhou, Siqi Zheng, Kai Hu, Ju Fan, and Chang Zhou. Contextual expressive text-to-speech. 2022.

Shubhi Tyagi, Marco Nicolis, Jonas Rohnke, Thomas Drugman, and Jaime Lorenzo-Trueba. Dynamic prosody generation for speech synthesis using linguistics-driven acoustic embedding selection. In *Interspeech 2020*. ISCA, oct 2020. doi: 10.21437/interspeech.2020-1411.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL `https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf`.

Michael Wagner and Duane G. Watson. Experimental and theoretical advances in prosody: A review. *Language and Cognitive Processes*, 25(7-9):905–945, may 2010. doi: 10.1080/01690961003589492.

Peilu Wang, Yao Qian, Frank K. Soong, Lei He, and Hai Zhao. Word embedding for recurrent neural network based TTS synthesis. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, apr 2015. doi: 10.1109/icassp.2015.7178898.

Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning*, pages 23318–23340. PMLR, 2022.

Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, Quoc Le, Yannis Agiomyrgiannakis, Rob Clark, and Rif A. Saurous. Tacotron: Towards end-to-end speech synthesis. March 2017.

Yuxuan Wang, Daisy Stanton, Yu Zhang, RJ-Skerry Ryan, Eric Battenberg, Joel Shor, Ying Xiao, Ye Jia, Fei Ren, and Rif A Saurous. Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis. In *International Conference on Machine Learning*, pages 5180–5189. PMLR, 2018.

Pengfei Wu, Junjie Pan, Chenchang Xu, Junhui Zhang, Lin Wu, Xiang Yin, and Zejun Ma. Cross-speaker emotion transfer based on speaker condition layer normalization and semi-supervised training in text-to-speech. *arXiv preprint arXiv:2110.04153*, 2021.

Yihan Wu, Xi Wang, Shaofei Zhang, Lei He, Ruihua Song, and Jian-Yun Nie. Self-supervised context-aware style representation for expressive speech synthesis. In *Interspeech 2022*. ISCA, sep 2022. doi: 10.21437/interspeech.2022-686.

Yujia Xiao, Lei He, Huaiping Ming, and Frank K. Soong. Improving prosody with linguistic and bert derived features in multi-speaker based mandarin chinese neural TTS. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, may 2020. doi: 10.1109/icassp40776.2020.9054337.

Hui Y. Xiong, Babak Alipanahi, Leo J. Lee, Hannes Bretschneider, Daniele Merico, Ryan K. C. Yuen, Yimin Hua, Serge Gueroussov, Hamed S. Najafabadi, Timothy R. Hughes, Quaid Morris, Yoseph Barash, Adrian R. Krainer, Nebojsa Jojic, Stephen W. Scherer, Benjamin J. Blencowe, and Brendan J. Frey. The human splicing code reveals new insights into the genetic determinants of disease. *Science*, 347(6218), jan 2015. doi: 10.1126/science.1254806.

Guanghui Xu, Wei Song, Zhengchen Zhang, Chao Zhang, Xiaodong He, and Bowen Zhou. Improving prosody modelling with cross-utterance bert embeddings for end-to-end speech synthesis. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, jun 2021. doi: 10.1109/icassp39728.2021.9414102.

Yi Xu. Speech prosody: A methodological review. *Journal of Speech Sciences*, 1(1): 85–115, 2011.

Dongchao Yang, Songxiang Liu, Rongjie Huang, Guangzhi Lei, Chao Weng, Helen Meng, and Dong Yu. Instructtts: Modelling expressive tts in discrete latent space with natural language style prompt. 2023. doi: 10.48550/ARXIV.2301.13662. URL `https://arxiv.org/abs/2301.13662`.

Zhenhui Ye, Zhou Zhao, Yi Ren, and Fei Wu. Syntaspeech: Syntax-aware generative adversarial text-to-speech. *IJCAI-2022*, April 2022.

Takayoshi Yoshimura, Keiichi Tokuda, Takashi Masuko, Takao Kobayashi, and Tadashi Kitamura. Simultaneous modeling of spectrum, pitch and duration in hmm-based speech synthesis. In *Sixth European Conference on Speech Communication and Technology*, 1999.

Chengzhu Yu, Heng Lu, Na Hu, Meng Yu, Chao Weng, Kun Xu, Peng Liu, Deyi Tuo, Shiyin Kang, Guangzhi Lei, Dan Su, and Dong Yu. DurIAN: Duration informed attention network for speech synthesis. In *Interspeech 2020*. ISCA, oct 2020. doi: 10.21437/interspeech.2020-2968.

Frank Zalkow, Prachi Govalkar, Meinard Müller, Emanuël A. P. Habets, and Christian Dittmar. Evaluating speech–phoneme alignment and its impact on neural text-to-speech synthesis. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, jun 2023. doi: 10.1109/icassp49357.2023.10097248.

Heiga Zen, Rob Clark, Ron J. Weiss, Viet Dang, Ye Jia, Yonghui Wu, Yu Zhang, and Zhifeng Chen. Libritts: A corpus derived from librispeech for text-to-speech. In *Interspeech*, 2019. URL `https://arxiv.org/abs/1904.02882`.

Guangyan Zhang, Ying Qin, Wenjie Zhang, Jialun Wu, Mei Li, Yutao Gai, Feijun Jiang, and Tan Lee. iemotts: Toward robust cross-speaker emotion transfer and control for speech synthesis based on disentanglement between prosody and timbre. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.

Ya-Jie Zhang and Zhen-Hua Ling. Extracting and predicting word-level style variations for speech synthesis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:1582–1593, 2021. doi: 10.1109/taslp.2021.3074757.

Kun Zhou, Berrak Sisman, Rui Liu, and Haizhou Li. Seen and unseen emotional style transfer for voice conversion with a new emotional speech dataset. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, jun 2021. doi: 10.1109/icassp39728.2021.9413391.

Kun Zhou, Berrak Sisman, Rui Liu, and Haizhou Li. Emotional voice conversion: Theory, databases and esd. *Speech Communication*, 137:1–18, 2022a. ISSN 0167-6393.

Yixuan Zhou, Changhe Song, Jingbei Li, Zhiyong Wu, Yanyao Bian, Dan Su, and Helen Meng. Enhancing word-level semantic representation via dependency structure for expressive text-to-speech synthesis. *Interspeech 2022*, April 2022b.

Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. A robustly optimized BERT pre-training approach with post-training. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1218–1227, Huhhot, China, August 2021. Chinese Information Processing Society of China. URL `https://aclanthology.org/2021.ccl-1.108`.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.